# Constraint-based modelling of metabolism in *Arabidopsis thaliana*

Alexander William Calderwood

A thesis submitted for the degree of
Doctor of Philosophy

University of East Anglia
John Innes Centre
September 2016

# Abstract

Plants are the most abundant biomass on Earth. Understanding plant metabolism represents a significant, fundamental challenge, requiring the incorporation of many fields of study. However it also provides potentially significant leverage with which to change the world in which we live.

The model organism *Arabidopsis thaliana* is probably the single best understood plant system. The aim of this thesis is to use mathematical modelling to investigate to what extent existing knowledge can describe broad, emergent aspects of the behaviour of metabolism in this system, with particular respect to the metabolism of sulfur, and other nutrients, and to gain insight into the consequences of the structure of its metabolic network.

Constraint-based modelling approaches provide a framework for modelling large reaction networks. Although they require various simplifications, and assumptions, they provide a route for the understanding of large metabolic networks, which is not possible through other approaches.

Here, a genome scale model of Arabidopsis metabolism is developed to reflect experimental data, and deployed in the study of nutrient stress, and nutrient requirements. This model predicts changes in gene expression in response to stress, and provides insight into the consequences of the metabolic structure on nutrient use efficiency, metabolic flexibility, and the consequences of genetic perturbation.

# Contents

# Associated publications

This thesis is associated with the published works below.

A. Koprivova, A. Calderwood, B-R. Lee, S. Kopriva, "Do PFT1 and HY5 interact in regulation of sulfate assimilation by light in Arabidopsis?", *Febs Letters*, 2014, doi:10.10.1016/j.febslet.2014.02.031

A. Calderwood, S. Kopriva, "Hydrogen sulfide in plants: from dissipation of excess sulfur to signalling molecule", *Nitric Oxide*, 2014, doi:10.1016/j.niox.2014.02.005

A. Calderwood, R. Morris, S. Kopriva, "Predictive sulfur metabolism — a field in flux", *Frontiers in Plant Science*, 2014, doi:10.3389/fpls.2014.00646

S. Kopriva, A. Calderwood, S. C. Weckopp, A. Koprivova, "Plant sulfur and big data", *Plant Science*, 2015, doi:10.1016/j.plantsci.2015.09.014

A. Calderwood, S. Kopriva, R. Morris, "Transcript abundance explains mRNA mobility data in Arabidopsis thaliana", *Plant Cell*, 2016, doi:10.1105/tpc.15.00956

# Acknowledgements

I would like to thank my supervisor, Richard Morris for his support throughout this project. He has provided a great sounding board for the discussion of ideas over the years, and in particular provided a number of important insights with regards to the analysis of mRNA. I would also like to thank him for his seemingly enormous patience during various stages of writing, and in particular redrafting, various things.

Thank you also to my other supervisor Stan Kopriva for providing me with the opportunity to study at the John Innes Centre, and providing enormous support, and encouragement, particularly in the early years of the project.

I also thank Joern Behre, who was a fantastically helpful collaborator, and who writes very nicely documented code.

Beyond the professional, thank you to Ben Hall, Daniel Knevitt, Lizzy Thursby, Matt Evans, Tom Vincent, & Izzy Web, who have provided some much needed perspective, and of course to Rachel Goddard who has by turns been inspirational, supportive, and mostly very nice.

Alex Calderwood
John Innes Centre, Norwich
September 2016

# Chapter 1

# Introduction

In this chapter we first introduce, and motivate the study of plant metabolic networks. We then describe the various analytical frameworks used to study metabolism, at various scales, in order to justify our choice of constraint-based modelling methods in the remainder of this work. Finally we discuss the creation of models for constraint-based modelling, and the strengths and weaknesses of constraint-based methods, with particular emphasis upon flux balance analysis, and elementary flux modes, the methods predominantly used in this thesis.

## 1.1 Why model plant metabolism?

Plants capture $121.7 \times 10^9$ metric tons of carbon per year from the atmosphere [15], from which they are estimated to produce $>$ 200,000 compounds. They therefore potentially have great scope for addressing issues of food, and fuel production, as well as the for production of interesting bioactive molecules.

However, plants have not generally achieved an anthropocentrically optimal phenotype. Crops require extensive fertilisation, the production of which is energy intensive [74], and potentially limited by diminishing resources [105], whilst elsewhere, high, toxic concentrations of mineral elements can inhibit growth, and lead to reduced yield [80]. Increasingly volatile environmental conditions, are likely to lead to relatively rapid changes in the ideal phenotypes of cultivated species, and a requirement for a widespread increase in resistance to biotic, and abiotic stresses. Furthermore, although secondary metabolites account for more than a third of all therapeutic compounds [150], they are generally produced in very small quantities in the host organism, and can often be more cheaply chemically synthesised.

Some of these targets may be accessible by conventional breeding, however, there is great interest in the engineering of plant metabolism. This is not without challenges, and early attempts to modulate the production of metabolites in the native organism often resulted in poor results, and/or off target effects [255, 26]. Technical challenges include the development of tools and methodologies for working with non-model organisms, (although methods continue to improve [23]), but also that metabolism in the native host is often tightly regulated so as to counteract any simple modification.

Abstracting beyond these difficulties, the focus of the approaches used here is determining which reaction steps should be modulated in order to bring about the desired result, assuming that any desired intervention can be achieved. By understanding metabolism, at the levels of the distribution of metabolic flux within an organism, and the distribution of control of this flux among reactions, a more rational approach to engineering can be achieved. The benefit of this design paradigm over one based purely on biological intuition can be seen in the growing number of successful microbial 'cell-factory' type studies (reviewed [58]).

In plants, metabolic engineering is full of promise, but success continues to be relatively rare [250]. This is generally attributed to their complexity in comparison to microbes. This occurs at all levels from gene regulation to physiognomy, but is normally discussed in an engineering context with particular reference to compartmentalisation at various levels [1]. Compartmentalisation both of specialised tissues and organs, and subcellular compartments leads to a heterogeneity, and a spatial aspect to metabolism, which is largely ignored in microbiology. There is significant experimental challenge to subcellular measurements, both of flux, and metabolite concentration, but additionally, knowledge of transporter steps, and enzyme localisation is generally poor relative to knowledge of enzyme function, even leading to uncertainty at the level of a biochemical reaction map. However, significant progress in both experimental [250, 1, 98], and analytical (discussed below) methods have led to the suggestion that 'rational' metabolic engineering is ready to begin to move beyond microbes, and into plants.

## 1.2 Overview of metabolic modelling approaches at different scales

Efforts to understand metabolism through modelling can be broadly split into two approaches: studies which use enzyme kinetics in order to derive flux control coefficients (FCCs) directly, using enzyme kinetic data, and efforts to predict the distribution of fluxes within the network, without considering kinetics. Predicted flux distributions can then be used to approximate control coefficients, and other properties of the metabolic network. Here we briefly discuss a range of modelling approaches at different scales, in order to contextualise, and justify

the approaches that we have taken in the body of this work.

### 1.2.1 Small models & the distribution of control

FCCs express the degree of control that an enzyme exerts over flux through some reaction of interest in the pathway (Figure 1.1). Ideally, FCCs can be derived for the system of interest, as these identify the level of control that each reaction has over the flux of interest, and therefore can be extremely useful for suggesting optimal metabolic interventions.

The distribution of control coefficients among the reactions of a pathway can be accessed through experiments in which an enzyme's activity is perturbed, and alterations to flux through the reaction of interest measured. However, the need to generate several lines with quantitatively different activities for each enzyme in the pathway to estimate its control coefficient, means that this is an arduous task, and in practice is rarely accomplished [231]. Furthermore, this approach is limited, as multiple controlling enzymes and non-linear dynamics make extrapolating behaviour away from measured conditions, (i.e. under different environments, and after genetic perturbation) difficult [184].

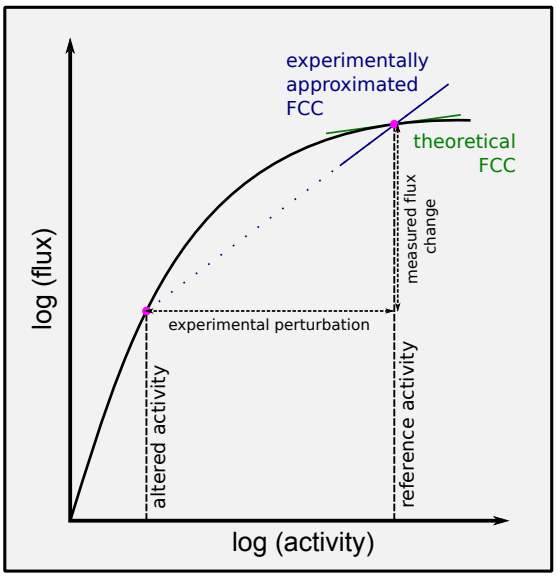



Figure 1.1: Flux control coefficients. Plot of log of test enzyme activity against flux through a reaction of interest to determine the control that the test enzyme exerts over flux through the reaction of interest. Flux control is the sensitivity of the flux through the reaction of interest in response to altered enzyme activity. Flux control can be experimentally estimated, but only crudely, a more complete understanding can often be gained through kinetic modelling.

Although experimental investigations into control of flux through the pathway are useful in qualitatively identifying important enzymes, kinetic models of the pathway can potentially be used not only to calculate control coefficients at the reference state more accurately than is possible experimentally (Figure 1.1), but also to simulate altered conditions. The difficulty however lies in producing an accurate model.

Kinetic modelling of metabolic pathways is well established (reviewed [39]), systems of ordinary non-linear, differential equations which specify the rate of reactions as functions of metabolite concentrations can be solved numerically using a range of freely available software [36]. This allows not only dissection of flux control distribution [39, 40, 141], but also predictions about how environmental perturbation changes control [141], and suggested engineering interventions to modify levels of metabolites [40]. However, as implied by the name, kinetic models require an extremely detailed understanding of the enzyme kinetics of the reactions in the studied pathway. This is the biggest hurdle to model building, particularly given that isoenzymes in different tissues or compartments often display different kinetics.

Strategies for determining parameter values can be broadly split into measurement, and estimation approaches. For small models, it may be possible to directly measure all kinetic parameters required [40]. However, the large experimental effort required [213, 226, 228] makes this a comparatively rare example; it is more common to search the literature to recover the majority of parameters required [177]. However generally poor coverage, particularly for allosteric regulation, means that it is accepted practice to use whichever parameters are available, either from experiments under differing conditions, or from orthologous proteins [177]. The validity of transferring parameters in this way is generally unclear [213], the exception being enzyme activity parameters, which are acknowledged to vary so greatly with environment, that they should be measured under the condition of interest [39].

Enzyme assay conditions used are typically far from the *in vivo* environment seen by the enzyme. Initiatives to design more '*in vivo* like' *in vitro* media, are underway for several microorganisms [62, 69, 124], but to the best of our knowledge, no such effort has been reported in plants, where the problem is exacerbated by the presence of multiple subcellular compartments, each with a unique environment.

The difficulty of obtaining experimentally measured kinetic parameters means that in the vast majority of published models, at least some parameters are fitted by minimising the difference between model predictions (e.g. of flux through the path), and experimental measurements [226]. One common problem with this approach is overfitting; assigning parameter values to fit the data more precisely than is justified. This can be seen, as many models parameterised using this top down approach lose predictive accuracy as conditions move away from those at which the parameters were fitted [81].

A minimal model of a subsystem should include everything that affects the internal variables of the model [39]. When considering broad issues, such as nutrient requirements, a much larger metabolic network should be considered when modelling nutrient assimilation than just the pathway itself. For example integration of sulfur assimilation within the wider metabolic network is demonstrated by the tight coordination of sulfur uptake with nitrogen and carbon availability [113, 112, 149]. Cysteine links both nitrogen and carbon metabolism to sulfur assimilation via O-acetyl-serine. O-acetyl-serine availability is a dominant factor in regulating the production of cysteine [19], and so its availability has to be considered in models of cysteine synthesis. This link to wider metabolism, and the ability to produce carbon-skeletons for nutrient integration has to be acknowledged when considering nutrient uptake and assimilation.

Unfortunately, as model size increases, the problems of unknown parameters, becomes extremely difficult to overcome, either by measurement, or estimation. To generate a large kinetic model, simplifying assumptions about parameter values [204], and rate laws [2], are frequently made, but this often results in poor model quality away from the fitted conditions [27] and so is of limited predictive value. Some specialised kinetic modelling approaches reflect structural, parameter, and rate law uncertainty in their predictions [53, 212, 222, 146]. These have resulted in the useful production of large kinetic models, with approximately 200 metabolites and reactions [106], but do not scale well to bigger models. As model size increases, parameter space expands enormously [251], resulting in prohibitive computational requirements [127]. As such, kinetic modelling currently does not scale to the size that is likely to be required to study nutrient uptake pathways.

### 1.2.2 Predicting flux vectors though constraint-based modelling

In contrast to the kinetic modelling approaches described above, constraint-based modelling provides a number of scalable, largely parameter free methods for understanding flux through large metabolic networks [125, 22]. As such, they are currently the methods which are best deployed to answer questions which must consider broad areas of metabolism, such as nutrient stress responses.

The only information reliably available across large regions of metabolism is often the structure of the reaction network itself, that is, which metabolites can be converted into which others. The absence of kinetic data means that it is impossible to apply a true metabolic control analysis to derive flux control coefficients as described above. However, by analysing only the network structure it is possible to predict flux distribution among reactions in the metabolic network. Furthermore, by analysis of the predicted flux distribution, and how it varies in response to changes in the network structure, (for example due to the removal of reactions), it is possible to approximate a control analysis of the

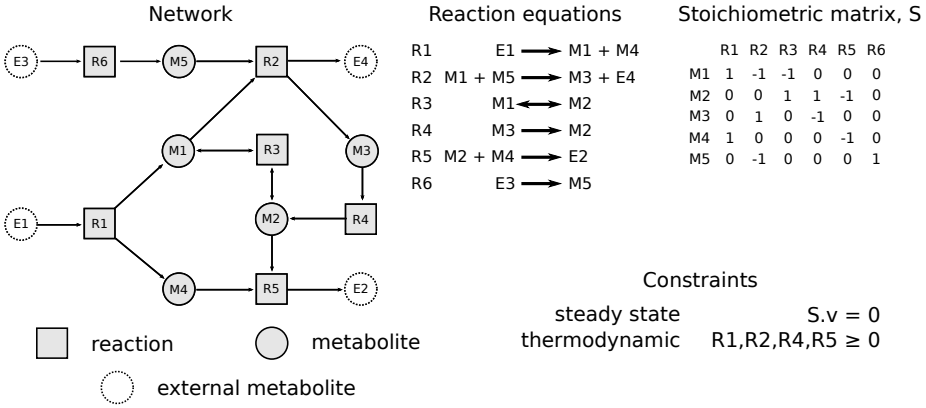



Figure 1.2: Example of the network flux vector problem for a toy network. The network itself, the associated reactions equations, and the stoichiometric matrix, which describes the metabolites consumed and produced by each reaction are shown. The aim of constraint-based methods is to solve the reaction flux vector $v$ for the constraints shown.

network, to simulate the effects of knocking out enzymes , and potentially to gain useful insights for metabolic engineering.

Figure 1.2 indicates the problem formulation prior to the application of constraint-based modelling methods. The 'reaction network' is a directed bi-partite graph, consisting of the metabolites and reactions present within the system. An edge between a metabolite and a reaction indicates that is is a substrate, or product of the reaction (indicated by edge directionality). Reactions which are considered to operate in both directions are indicated by double headed arrows. Although not exemplified in Figure 1.2, edges are weighted to reflect reaction stoichiometry, however, generally no kinetic information is considered.

Internal metabolites are those entirely within the system, whereas external metabolites are those which are considered to be exchanged with the surrounding environment. At steady state, there is no accumulation of internal metabolites in the system. Therefore for each internal metabolite, flux through the reactions which consume and produce it must be equal. This constraint is not applied to external metabolites, which are considered to exist in great excess in the surrounding environment. Reactions which transform metabolites within the system are considered internal reactions, while reactions which transport metabolites in and out of the considered system are here called 'exchange reactions'. Due to thermodynamic considerations, some reactions are effectively irreversible under physiological conditions, and must proceed in an appropriate direction. Depending on the analytical method used, these reactions are bounded so as to only carry flux in the positive direction ($v_i \geq 0 \quad \forall\, i \in$ irreversible reactions), or reversible reactions may be split into forward and reverse components in the stoichiometric matrix, and all elements of $v$ be greater than or equal to 0.

By assuming that the system exists at steady state, the problem can be expressed as

$$S \cdot v = 0, \tag{1.1}$$

where $S$ is the stoichiometric matrix of the reaction network, which contains information about which reactions produce and consume each metabolite, and $v$ is the flux vector, of the flux through each reaction in the network. As will be discussed below, this problem statement allows the application of various mathematical approaches to solving for $v$ within a biological context. However, first we will discuss the creation of the reaction network graph in greater detail.

## 1.3 Building constraint-based models

The quality of output of all constraint-based methods depends on the quality of the reaction network, and how similar it is to the biological system. The construction of a genome scale model is not a facile task, and particularly in plants remains a laborious undertaking [182, 201, 218]. Briefly, allowed metabolic reactions are recovered from an annotated genome. The stoichiometry of reactions must then be checked, to ensure mass-balance, and as far as possible reactions assigned to compartments, based on enzyme location information. A 'biomass equation', representing the drain of some set of metabolites to produce biomass is then added, as well as pseudo-reactions for growth associated energy requirements in the form of ATP and reducing agents. Thermodynamic constraints for reaction reversibility must be added. Finally the predictions of the model should be compared to experimental data, and the structure of the model updated as required. This process is expected to take from weeks, to over a year depending on organism complexity [218].

Some of the above steps can be automated [102, 49], accelerating the process, to approximately 1 week for production of a draft model from genome sequence [49], however many steps require extensive manual curation and validation against the literature. The quality of model predictions generally correlate to the amount of manual curation carried out, as reflected in the relative quality of the labour intensive, manually constructed models of *Arabidopsis thaliana* metabolism [5, 33] in comparison to other approaches.

As sequencing technology improves more and more species are routinely sequenced. This has enabled the creation of many genome scale databases of reactions and pathways in different systems. The BioCyc database collection (`biocyc.org`, [253]) hosts a continuously extended and updated collection of databases from 7,615 species including a wide variety of plants. These databases are divided into three tiers based on quality; tier 1 databases have each required at least one person year of manual curation, whereas tier 2 & 3 contain computationally predicted metabolic pathways based on genome sequencing, and

homology to enzymes in other species, tier 2 databases have additionally received some manual curation. Although a diverse number of plant species are included in these databases, Arabidopsis is the only plant species with a tier 1 database.

These databases provide a strong foundation for constraint-based models, but they cannot be used directly, as generally they are incomplete, with no possible steady state solution including flux through a reasonable biomass equation. Deciding which additional (relatively unevidenced), reactions to include to resolve this remains a somewhat subjective, time consuming process.

Beyond modelling metabolism alone, there have been some initial approaches to integrate constraint-based models as one level in a multi-scale model. This commonly takes the form of overlaying a metabolic model with boolean, or petri-net models of regulatory, or signalling processes. These then modify reaction constraints within the metabolic model in response to their output [37, 114, 59]. The output of the constraint-based metabolic model can then be fed back into the model of regulation, and the whole system iteratively updated. Similar hierarchical modelling approaches can also be used to embed constraint-based models of particular tissues within empirical, whole body scale kinetic models of the organism [71], to provide particularly fine-grained mechanistic detail about particular aspects of the broader model, or conversely to embed some kinetic details to modify the flux boundary constraints in the constraint-based mode [38]. In chapter 2, we use small embedded kinetic models to modify uptake flux boundaries in response to sulfur starvation.

### 1.3.1 Difficulties associated with plants

As ever, plants prove to be trickier than microbes, for a number of reasons. Here we discuss these difficulties as they apply to building constraint-based models of metabolism.

#### 1.3.1.1 Compartmentalisation

Spatial compartmentalisation of reactions is incorporated in the reaction network by the duplication of reactions and metabolites for the different compartments they are found in. Compartment information is normally appended to the name of metabolites and reactions. A reaction in a given compartment is only able to produce and consume metabolites located within the same compartment. Transport of metabolites between compartments is represented by a reaction in the same way that enzymatic conversions are. For example, the transport of metabolite $A$ from compartment $a$ to compartment $b$ is represented by

$$A_a \xrightarrow{transport} A_b,$$

as compared to the conversion of metabolite $A$ into metabolite $B$ within compartment $a$, which is represented as

$$A_a \xrightarrow{conversion} B_a.$$

Although compartmentalisation can be easily incorporated within the constraint-based network, experimental uncertainty as to which enzymes and metabolites are located in which compartments, and which metabolites can be transported between compartments leads to errors in the model building, as it is not necessarily clear which reactions have access to which metabolite pools. Duplication of reactions across compartments also leads to larger models, which exacerbate the computational difficulties of some analytical frameworks. Additionally, multiple occurrences of reactions can often act in a (partially) compensatory manner, making it difficult to distinguish the flux through each compartmentalised instance of the reaction.

Plant cells probably exhibit the greatest metabolic complexity of all living organisms due to their extreme level of subcellular compartmentalisation. Additionally, key metabolic pathways overarch multiple compartments. For example the photorespiratory carbon oxidation pathway requires 12 transporter reactions to link all involved compartments involved together for all relevant metabolites [33].

Most published genome scale models of plant metabolism include compartmentalisation to some extent, but the quality and coverage of these assignations is not always clear. Wide variation between models in which compartments reactions occur [169, 139, 145] suggests that the number of reactions which are assigned to particular compartments, and in particular to the vacuole, is probably much lower than in reality [115]. This reflects the current difficulty of experimentally determining sub-cellular reaction location.

#### 1.3.1.2 Secondary metabolism

Primary metabolism is relatively well understood, however the generally poorer understanding of the vast plant secondary metabolism, is reflected in the focus of models published to date [5, 33, 169, 244]. The large number of unknown metabolites was recently highlighted for sulfur metabolites in particular [68], and the potentially missing reactions suggested by the proportion of genome content with unknown function [196] could also adversely affect prediction quality.

Although large parts of plant metabolism can be expected to be missing, or poorly incorporated in many models, this does not necessarily undermine the predictions made. Predicted fluxes through central metabolism have been found to be accurate in comparison to experimental data [33], and the quality of predictions can therefore be expected to depend to some extent on the area of metabolism being considered. A related difficulty is that $^{13}$C tracer experiments

typically used for model validation can generally only be used to determine flux through central metabolism, and it can be therefore be difficult to validate model accuracy for secondary metabolism, and therefore to improve the quality of these regions of the reaction network.

#### 1.3.1.3 The biomass equation

The stoichiometry of the biomass equation is normally based on experimental assessments of the composition of the organism of interest. However, these experimental methods are unable to identify all metabolites required for biomass production, and, for example, relatively scarce cofactors are normally not included. This also highlights the somewhat tricky and subjective distinction between 'biomass precursors', which are somehow considered to be an 'end product' of the plant metabolism, as compared to other metabolites, which merely facilitate their production.

In plants, tissue and cellular composition varies widely, not only between between different tissues, but also under different environmental conditions (e.g. [111]). Accurate tissue specific measurements under the conditions of interest are therefore required, but difficult to achieve practically. Plants are also known to lose a large proportion of photosynthate into the soil, which as far as we are aware has not been considered in determining the biomass equation in any study.

#### 1.3.1.4 Tissue specific models

It seems increasingly unlikely that all known reactions catalysed by enzymes encoded in the genome of an organism are utilised, either within a given tissue, or cell type, or within a particular physiological environment. For example, certain metabolic pathways are unique to particular tissues, e.g. the light and dark photosynthesis reactions differ between green tissues under an autotrophic lifestyle, and other tissues, but also within green tissues during the day and night. It may be of interest to produce specific submodels, more closely reflecting these particular circumstances.

All tissues within the plant contain the same genetic information, all the time, and therefore could potentially produce the same enzyme expression patterns. Two approaches can therefore be used to capture these tissue specific differences: either by working out what the flux distribution is attempting to achieve in a particular tissue, and assuming that gene expression regulation is such that this is achieved [31] (the validity of this is discussed below), or by imposing additional constraints as to which subset of reactions in the model are permitted to carry flux in a tissue specific submodel [134].

Many methods exist for the integration of tissue specific experimental data to constrain flux solutions, and make tissue specific submodels. Although often

there is not complete information about where every reaction occurs, it is common to approximate this using tissue specific transcription, or protein expression data [182, 145]. A large number of related approaches for incorporating these data exist (recently reviewed [233]). These can be broadly split into two strategies. In various 'switch & valve' approaches, different thresholding approaches are used to constrain the extent to which reactions are permitted to carry flux, based on the expression of their associated genes. Conversely, in optimisation approaches, genes are classified as either desirable, or not, based on transcript expression, or other experimental data. Solutions are either found which maximise the occurrence of desirable reactions, or which minimise the number of undesirable reactions, without compromising the ability of desired reactions to carry flux [188]. However, the application of these methods to the same models and datasets often yield very different submodels [52], and it is not clear that the quality of model predictions are improved by this use of experimental data to provide additional constraints. Machado et al. [130], assessed the quality of seven methods for integration of transcriptomic data to constrain the predicted flux distribution, and found that none of them consistently outperformed simulations which completely ignored transcriptomic data.

Given the relative ease with which tissue specific models can be made, a natural next step is to link various tissue models together. This can be easily achieved simply by allowing reaction networks for the specific tissues to interact at particular metabolites [31, 44]. The difficulty lies in identifying the metabolites which they should be permitted to exchange, which depending on the modelled system can be difficult experimentally, and which can be expected to dramatically affect model results. Unusually, many pioneering examples of this multi tissue approach have been achieved in plants. Dal'Molin et al. [43] produced a model of the interaction between the bundle-sheath, and mesophyll cells in $C_4$ metabolism, Cheung et al. [31], modelled the interaction of autotrophic cells in the day, and night through storage molecules, Bogart & Myers built a contraint based model of a maize leaf [21], and recently, Dal'Molin et al. [44] investigated the translocation costs associates with spatial separation of biosynthetic activities in a multi-tissue model.

## 1.4 Methods for constraint-based modelling

Once a stoichiometric model of metabolism is created, constraint-based modelling approaches assume that the system is at metabolic steady state, such that no internal metabolites undergo net production or consumption, in order to deduce information about the possible flux through each reaction in the system.

The steady state assumption is expressed as $S \cdot v = 0$, where $S$ is the stoichiometric matrix for the reaction network, and $v$ is the flux vector to be determined. Stoichiometric metabolic flux balance analysis is the simplest approach to de-

ducing the flux vector. As illustrated in i-iii of Figure 1.3a, linear algebra can be used to find a flux distribution which follows the steady state constraint (ii). Additionally relatively easily measured uptake fluxes of metabolites into the system can further constrain the flux solution (iii). A flux solution can be iteratively obtained to best fit the data, by minimising the sum of squared differences,

$$\text{minimise} \sum \frac{(v - v_m)^2}{\sigma_v^2},$$

between the predicted and experimental uptake flux data, weighted by uncertainty in the experimental data.

This approach is not computationally or experimentally demanding. However, the problem is generally underdetermined by the constraints of network stoichiometry, and measured uptake fluxes, meaning that is insufficient to determine all internal fluxes (as demonstrated in Figure 1.3aiii). To resolve this problem the considered pathway can potentially be simplified through the removal of reactions which are assumed to carry no, or little flux. However, the simplification required generally is too drastic to allow consideration of the question of interest.

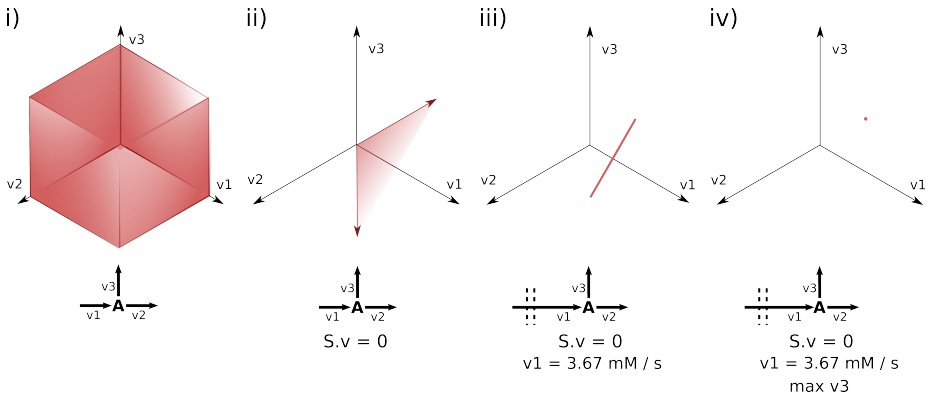
Beyond system simplification, several different approaches can be taken to deal with this problem: additional experimental data can be collected to further constrain the flux distribution ($^{13}$C-metabolic flux analysis,$^{13}$C-MFA), additional modelling assumptions can be made (flux balance analysis, see FBA), or the properties of the underdetermined space itself can be explored (elementary flux mode (EFM) analysis).

### 1.4.1 $^{13}$C-MFA

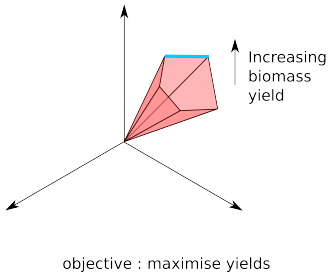
By incubation of cell suspensions with a labelled chemical, for example [1-$^{13}$C] glucose, $^{13}$C is taken up into the system, and generates a distinct accumulation of isotopomers, depending on flux through the reactions of the system. The concentration of isotopomers with distinct labelling patterns can be detected experimentally by NMR, or MS approaches, and, (in conjunction with an understanding of the atomic rearrangements involved in each reaction), used as additional constraints to the flux solution of the system.

An analytical solution to the flux distribution given isotopomer abundance data is available only for the simplest systems, and an iterative process of flux solution generation, and variation, is normally followed to determine flux solutions. This is achieved by finding the solution which minimises the sum of squares error to the experimental data such that

$$\text{minimise} \sum \frac{(x - x_m)^2}{\sigma_x^2} + \sum \frac{(v - v_m)^2}{\sigma_v^2}$$




(a) Toy example of flux balance analysis. A simple network consists of three reactions in which the metabolite, A, is produced and consumed. All reactions are considered to be unidirectional. i) The unconstrained flux distribution, all positive flux values are available, for each reaction. ii) The steady state assumption restricts permissible flux space to the plane v1 = v2 + v3, iii) if in addition, uptake flux, v1, is measured (and in this case found to be 3.67mM/s), the possible flux distributions are restricted to the line 3.67 = v2 + v3. iv) in the defining step of flux balance analysis, we assume that the metabolic network is regulated so as best to perform some metabolic function. In this example we assume that the objective function is to maximise flux through v3, giving the unique solution v1 = 3.67 mM/s, v2 = 0 mM/s, v3 = 3.67 mM/s.




(b) In most real world applications of flux balance analysis, the assumption of optimality does not lead to a unique solution. In this fictitious example, the objective function (as is commonly used), is maximisation of the yield of biomass, which leads to degenerate optimal solutions, along the highlighted edge. Approaches to deal with this problem include a secondary optimisation step, in which a second independent objective function is imposed in order to further reduce the solution space, FVA, which returns the maximum, and minimum flux through each reaction under the optimality criterion, or one of a variety of approaches for sampling of the optimal (and occasionally slightly suboptimal) solution space.

Figure 1.3: Flux balance analysis to reduce permissible flux space.

where $x_m$ is the measured label, and $v_m$ are the measured uptake fluxes. $^{13}$C-MFA allows the consideration of larger, and more complex metabolic networks, than stoichiometric MFA alone, however it requires reasonable experimental effort, and can only consider metabolic systems proximal to the fed metabolite, and therefore can only be used in the study of particular aspects of metabolism.

Conventional $^{13}$C-MFA assumes metabolic, and isotopomer steady state, such that fluxes and isotopomer level are assumed constant over the experimental period. Extensions to this framework exist which do not require these assumptions [122, 97], and which allow the method to be extended to consider for example larger, or more complex metabolic systems through the use of multiple metabolite labels in parallel [123], and more efficient computational procedures [3].

Ultimately, MFA cannot be used to study large, genome scale networks. It requires extensive experimental effort and expertise, and provides a descriptive, rather than predictive approach to determining fluxes. It therefore cannot be used to predict the effect of for example genetic, or environmental perturbation. In this study, we focused on the other two approaches for dealing with an underdetermined system.

### 1.4.2 FBA

FBA [230] is a powerful technique to estimate internal flux distributions. As shown in Figure 1.3a, in addition to the steady states and thermodynamic constraints, FBA imposes some assumed objective function, usually in conjunction with a measured substrate uptake flux, in order to further reduce the flux solution space. This approach relies only on very limited experimental data, and can be formulated as a linear programming problem. Efficient algorithms, and sophisticated optimisation software for solving these kinds of problems means that optimal solutions can be calculated extremely rapidly, and FBA can be applied with impunity to the largest metabolic models.

FBA based methods are probably the most commonly used of the discussed approaches, and have been used for understanding metabolic efficiency [30], interpreting 'omics data [220, 202], predicting novel metabolic pathways [82], and how flux distribution changes in response to genetic and environmental changes [202, 197, 31]. However it is not without issues.

#### 1.4.2.1 Degeneracy

Traditional FBA produces only a single point estimate for flux through the system, although depending on the network structure, and objective function used there are likely to be a number of degenerate optimal solutions (Figure 1.3b). It is quite common to assume a secondary objective function, such as minimisation

of total flux, or minimisation of discrepancy between predicted flux distribution, and experimental data (typically transcriptomic) [186] in order to further reduce the solution space, however this is not guaranteed to result in a single solution. Although the degeneracy of predicted optimal distributions is often considered undesirable [172], it is in fact likely to reflect biological reality: degenerate optimal solutions are consistent with robustness, and a population of cells is unlikely to be adequately described by a single flux distribution [174, 118].

In order to more accurately represent the degenerate optimal solution space, after initial optimisation, flux variability analysis (FVA) [132, 205], applies this optimality as an additional constraint, before a second level of optimisation, in which flux through each reaction is sequentially maximised and minimised, in order to estimate the boundaries of the solution space. Flux variability is probably currently the most widely used FBA based approach. Unfortunately, whilst this method does define the hyperrectangle containing the optimal solution, it does not allow relationships between reactions within the optimal space to be determined. For example, in the simple network structure shown in Figure 1.4, FVA would determine that flux through reactions B & C are between 0 and 10 a.u., but not that B + C = 10 a.u. In order to gain a more sophisticated description of the optimal solution space, Cheung et al. [32] applied random preference weights to reactions in a secondary optimisation step in order to sample many solutions from within the solution space, which were then analysed as a group. Although this approach is well suited to the study of relatively small models, it is not clear how generally practical it is for the analysis of large populations due to the "curse of dimensionality" [79], and the huge number of samples that could potentially be required to adequately sample the solution space. As discussed below, elementary flux modes potentially offer a more thorough and concise description of the optimal flux space.

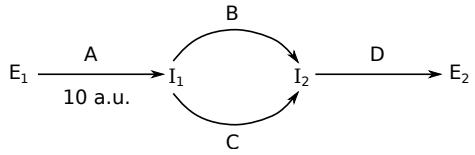



Figure 1.4: Example demonstrating the shortcomings of FVA. E represents external metabolites, and I, internal metabolites, which must be flux balanced at steady state. Flux through reaction A has been determined to be 10 arbitrary units. Although FVA can determine that the range of flux solutions for reactions B & C are both from 0 to 10 a.u., it cannot determine the relationship between them: that B + C = 10 a.u.

#### 1.4.2.2 The objective function

The striking weakness of the FBA approach is the pleasing but unproven assumption that metabolic networks have evolved so as to be regulated such that the flux distribution fulfils an objective function.

Even under this assumption, identifying what the objective may be is not trivial. A number of objective functions have been considered in the literature, most often tied either explicitly or implicitly to metabolic efficiency [30], although other objectives have been proposed which either aim to simulate, and maximise growth rate [252] or minimise conflict with 'omics data [14, 34]. In bacteria, the quality of fluxes predicted assuming maximisation of biomass varies with environment, growth phase, and species [192, 56], suggesting that the appropriate objective function is specific to physiological condition.

It is, however, not clear that organisms do necessarily act to optimise a single objective function [121]. Fischer & Stauer [57], demonstrated that a number of genetic mutants in "regulators of not-yet activated adaptive responses" exhibited improved biomass production relative to wild-type *B. subtilis*, and Ibarra et al. [94], demonstrated that *E. coli* grown under constant environmental conditions evolves towards a flux distribution consistent with maximum biomass production. Both of these studies indicate that flux in the wild-type form of these organisms is not distributed so as to maximise a single objective. Furthermore, experimentally measured fluxes in bacteria often exist in apparently suboptimal regions, which allow large variation in flux through individual reactions [78, 179] without further compromising the single assumed objectives. The extent to which apparent sub-optimal distributions arise through the averaging of measured fluxes in a heterogeneous population, rather than 'sub-optimality' in a single cell is unclear. Nevertheless, it seems likely that in multicellular organisms, in which cells are differentiated, the situation can be expected to be more complex still.

It is currently unclear whether 'suboptimal' flux distributions exist on a pareto optimality front, in which any alteration in flux distribution leads to decrease in at least one of the objectives [187], or simply in some suboptimal space, through incomplete, or otherwise noisy evolutionary processes. However, it has been argued that metabolism (in bacteria) exists on the tradeoff front between growth rate, and robustness to environmental perturbation [109]. Over thousands of generations of growth under constant, laboratory conditions, microorganisms can be expected to have been selected towards achieving maximal growth rates. It is interesting to speculate that this simplification of objectives might contribute to the relatively strong performance of FBA on cultured microorganisms, in comparison to, for example, plants.

FBA based methods are beginning to appear that address partially optimised distributions, and multiple objectives [245], but these remain a major challenge for the FBA framework. Given the sophistication and size of plant metabolic networks, and numerous differentiated cell types, they are likely to be particularly relevant to their study.

Pragmatically, although the use of an objective function is somewhat problematic, an increasing number of studies have accurately predicted flux distributions in plants cells using FBA (for example [33, 244, 83]). Additionally, in Arabidopsis, central carbon metabolism has been shown to be fairly insensitive to the

objective function used [33]. However, it is likely that this robustness depends both on the organism, and the particular area of metabolic interest, and different areas of metabolism can potentially be expected to behave differently.

### 1.4.3 EFM analysis

Elementary modes are minimal sets of reactions that can operate at steady state, with all reactions proceeding in thermodynamically feasible directions [190]. These sets are minimal, in the sense that no subset of reactions in an elementary mode is sufficient to carry flux at steady state. Figure 1.5 illustrates the EFMs of a simple toy network.

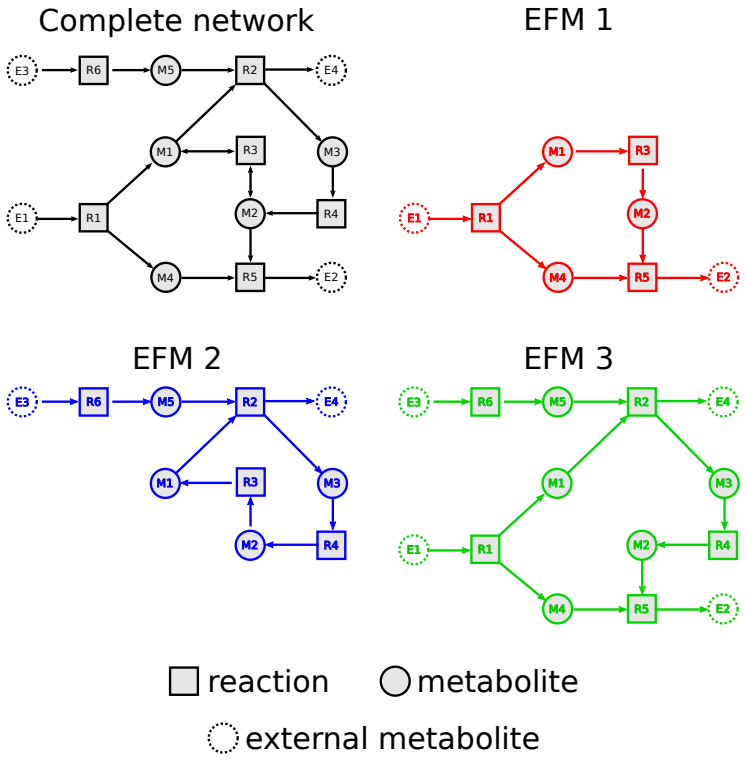



Figure 1.5: Elementary modes of an illustrative toy network. Each elementary mode is a minimal steady state solution, in that the removal of any reaction prevents flux through the whole of the mode. All possible steady state flux distributions in the complete network can be expressed by the superimposition of combinations of EFM1, 2, & 3.

Elementary modes are the edges of the flux-cone shown in Figure 1.3b, which pass through the origin. Therefore the span of elementary modes is all possible metabolic states, in that all possible steady state flux distributions can be expressed as a weighted sum of the elementary flux modes. Elementary modes

therefore allow a (relatively) compact description of the infinite possible flux states in the cell, and therefore its metabolic capabilities.

The advantage of elementary modes over the extremely similar 'extreme-pathways' concept [185], is that bi-directional reactions are decomposed into two uni-directional reactions, and that therefore every flux mode can be decomposed into EFMs without the need to consider potential 'cancelation' of flux through a reaction in opposite directions.

#### 1.4.3.1 Applications of elementary modes

Elementary modes have various applications associated with exploring the metabolic capabilities of the organism. Examination of EFMs allows, for example, determination of all minimal media for the growth of microorganisms, and discovery of alternative metabolic pathways [191, 167]. Yield space analysis [207], projects elementary modes onto a 2-dimensional surface, and the metabolic capability of the organism in these dimensions is encompassed by the convex hull of the plotted points. This allows the relationship between two fluxes to be inspected visually, typically the yield tradeoff between two products, for example a high value metabolite, and biomass. In chapter 3, we essentially use yield space analysis to examine the relationships between various nutrient requirements imposed by the constraints of the metabolic network.

Elementary modes can also be used to investigate, and understand transcriptomic, or fluxomic data, and to help interpret what the flux is "for" [171]. A number of methods have been proposed for decomposing a given experimentally determined flux distribution, either directly [171], or as implied by transcriptomic data [176, 95] into component elementary modes. Additionally, Stelling et al. [211], found that the 'control-effective flux' (CEF), of a reaction towards a given metabolic objective, could be calculated from EFMs, and correlated with mRNA expression data. The CEF of a reaction was calculated as the average flux through that reaction across modes, weighted by the efficiency of the EFM (in the production of biomass). The CEF of a reaction is related to the robustness of the system to perturbation of that reaction. Interestingly Stelling et al. [211] found that by calculating this metric across all elementary modes, rather than just the 'optimal' ones, the strength of correlation between CEF and mRNA expression increased, again suggesting the importance of considering sub-optimal metabolic modes in biological systems.

##### Metabolic robustness

This early interest in the use of EFMs for determining the robustness of a metabolic network has been extended in several studies. Although robustness is commonly defined as the insensitivity of a system to changes in external (environmental), or internal (genetic) parameters, the way that it is calculated

depends on the modelling framework being used, and topological, FBA, and EFM based metrics have been developed [121]. constraint-based measures for assessing robustness include the FBA based MOMA [197], and ROOM [200], which assess the minimal metabolic perturbation enforced by reaction removal. However, the results of both of these methods depend heavily on the assumed objective function of the system [121].

The simplest EFM-based measure used to assess robustness is simply the number of EFMs through the system [173], however this is generally regarded as inadequate. Wilhelm et al. [242], used a metric based on the average number of feasible EFMs remaining after a reaction is knocked out, an idea extended by Behre et al. [16], to consider multiple knockouts. However, the combinatorial explosion in knockout sets with size limits the extrapolation of this approach, as not all possible numbers of deletions can be considered by brute force. Furthermore, it relies of the calculation of all EFMs in the network. This metric has been shown to be mathematically equivalent to a more computationally tractable approach based on minimal cut sets, calculable without the need for all EFMs, [234, 65], allowing its approximate application to genome scale models. Another, similar approach is used by Min et al. [144], except that whereas Wilhelm and Behre consider all reactions as equally likely to fail, Min et al. weight the probability of knockout by reaction involvement across EFMs. This demonstrated that although metabolic networks can generally be considered robust to random perturbation, they may be fragile towards non-random, targeted intervention. This naturally leads to a consideration of EFMs for guidance of metabolic engineering strategies.

#### Metabolic engineering

Elementary modes have been used for the rational design of bacterial strains which overproduce particular metabolites of interest (reviewed [89]). Most approaches to EFM guided engineering centre around identifying desirable (efficient) modes for the production of the metabolite of interest, followed by the elimination of all inefficient modes [223, 224], by knocking out reactions involved only in inefficient modes. Methods for determining constrained minimal cut sets, the fewest interventions required to prevent inefficient modes, without compromising desirable ones have been developed [76]. One problem with these approaches is that although EFMs can identify the most efficient modes, the absence of enzyme kinetics means that they can not identify the modes which lead to the greatest overall production. The CASOP heuristic [77] tries to account for this by weighting reaction removal by EFM membership production efficiency, but allowing the retention of less efficient modes as well.

These cut-set methods remove inefficient modes, but do not allow for the discovery of the beneficial overexpression of reactions. FluxDesign [140] identifies reaction candidates for overexpression based on positive correlation between reaction flux, and flux to the product of interest across elementary modes. We

use an extremely similar, correlation based approach, to identify reactions important in determining nutrient requirements in chapter 3.

Application of these approaches have led to greater understanding, and to various positive interventions for the production of metabolites in bacteria, but have not been widely applied to other organisms. Although EFMs allow a complete understanding of the capabilities of a metabolic network, their number undergoes a combinatorial explosion as the size of the considered network increases, meaning that their use has largely been limited to only relatively small networks of tens, or hundreds of reactions. Most modern metabolic models consist of thousands of reactions, particularly in eukaryotes, where compartmentalisation tends to lead to reaction duplication. Consequently, the application of EFM analysis to plants has been generally limited to exploring the metabolic capabilities and interactions of small subsets of paths [167, 18].

#### 1.4.3.2 Calculation of elementary modes

We have discussed a number of applications of elementary modes, which offer the best current framework for understanding the metabolic capabilities of a system, but their application to date has been limited. This is because the number of elementary modes undergoes a combinatorial explosion with network size [110], leading to difficulties in their calculation for large scale networks, which can be trivially analysed by FBA approaches. A number of distinct methods have been developed for the calculation of elementary modes, but can be broadly divided into double-description based methods, which calculate elementary modes from the null-space of the stoichiometric matrix, and mixed-integer linear programming (MILP) approaches, for their calculation based on the non-decomposability criterion.

We will discuss both approaches in greater detail, however, the double-description based methods, are generally much faster, (able to practically generate millions of EFMs), but they do not iteratively calculate complete EFMs, and therefore must be able to calculate all EFMs for the network. Consequently they can generally only be applied to medium scale networks [235, 217, 100]. Conversely MILP-based methods are relatively slow, but iteratively return true EFMs, and so can be used to return subsets of the complete EFM set of the system, allowing their application to models of arbitrary size.

Although not widely used, and so not discussed here, 'conversion analysis' is an interesting approach which can similarly be used to either return EFM subsets of interest, or for particular reaction sets of interest [227].

### Double-description approaches

The null-space of a matrix, $S$, is the set of all solutions, $v$, such that

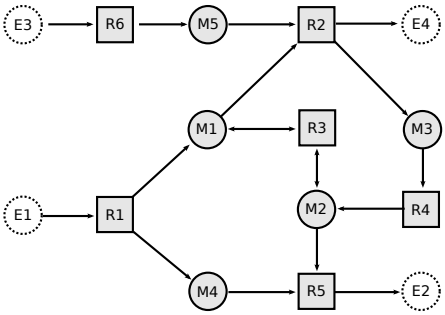
$$S \cdot v = 0.$$

When $S$ is the stoichiometric matrix of a reaction network (see Figure 1.2 for example of stoichiometric matrix and reaction network), the null-space is the set of all reaction flux vectors which result in steady state. A basis of the null-space can be easily, and rapidly calculated, such that the full null-space can be expressed as linear combination of columns of the basis. This basis is related to elementary modes, however it is not guaranteed to be biologically interpretable, as it may include negative (flux) values through reactions which can only operate in one (positive) direction. Double-description methods work by converting the found basis to a basis with only positive fluxes.

For an illustrative example of the basic double-description based approach see Figure 1.6. Essentially, working from the top of the calculated null-space basis, for each row (reaction), these methods combine any column with a negative element in the current row with all columns with a positive element, so as to cancel the current row to zero. This leads to a combinatorial explosion in the size of the working matrix, and the computational resources required. When the last row is reached, the full set of elementary modes is given, such that each column is an elementary mode. However, until the last row is reached, no column is necessarily a true EFM. This approach, together with various refinements to improve performance [61], has been implemented in widely used software tools [235, 217] and recently has been efficiently implemented, leading to an improvement in performance by several orders of magnitude, and is currently the fastest approach for EFM calculation [229].

A 'demand based network splitting approach' [93] can be used to specify only the enumeration of EFMs which either include, or don't include a given reaction. This approach, recursively applied, allows the parallelisation of calculation of EFMs using double-description based methods. Using this strategy, in conjunction with their improved software, van Klinken et al. [229] were able to calculate all elementary modes for a model consisting of 318 reactions, and 335 metabolites. Whilst this is an impressive technical achievement, it also illustrates the current challenge of calculating the complete set of EFMs for large genome scale models.

Beyond their simple calculation, the explosion in the number of EFMs also leads to difficulty with their storage and analysis. In their enumeration of the EFMs for the same model, Hunt et al. [93] report that EFM storage required ∼1TB of storage, and that "most of the elapsed time involved reading, decompressing, compressing, and writing the results". These difficulties have led to an interest in the calculation of particular subsets of EFMs rather than the full set. There are some methods for subset calculation using double-description based methods [131]. However, the majority of sampling approaches are based on a MILP




Stoichiometric matrix, S

| | R1 | R2 | R3 | R4 | R5 | R6 |
|---|---|---|---|---|---|---|
| M1 | 1 | -1 | -1 | 0 | 0 | 0 |
| M2 | 0 | 0 | 1 | 1 | -1 | 0 |
| M3 | 0 | 1 | 0 | -1 | 0 | 0 |
| M4 | 1 | 0 | 0 | 0 | -1 | 0 |
| M5 | 0 | -1 | 0 | 0 | 0 | 1 |

Irreversible stoichiometric matrix, S

| | R1 | R2 | R3f | R3r | R4 | R5 | R6 |
|---|---|---|---|---|---|---|---|
| M1 | 1 | -1 | -1 | 1 | 0 | 0 | 0 |
| M2 | 0 | 0 | 1 | -1 | 1 | -1 | 0 |
| M3 | 0 | 1 | 0 | 0 | -1 | 0 | 0 |
| M4 | 1 | 0 | 0 | 0 | 0 | -1 | 0 |
| M5 | 0 | -1 | 0 | 0 | 0 | 0 | 1 |

R1,R2,R4,R5 ≥ 0

Null-space

| | $N_1$ | $N_2$ | $N_3$ |
|---|---|---|---|
| R1 | 0 | 1 | 0 |
| R2 | 0 | 0 | 1 |
| R3f | 1 | 1 | -1 |
| R3r | 1 | 0 | 0 |
| R4 | 0 | 0 | 1 |
| R5 | 0 | 1 | 0 |
| R6 | 0 | 0 | 1 |

| | $N_1$ | $N_2$ | $N_{1,3}$ | $N_{2,3}$ |
|---|---|---|---|---|
| R1 | 0 | 1 | 0 | 1 |
| R2 | 0 | 0 | 1 | 1 |
| R3f | 1 | 1 | 0 | 0 |
| R3r | 1 | 0 | 1 | 0 |
| R4 | 0 | 0 | 1 | 1 |
| R5 | 0 | 1 | 0 | 1 |
| R6 | 0 | 0 | 1 | 1 |

| | $N_2$ | $N_{1,3}$ | $N_{2,3}$ |
|---|---|---|---|
| R1 | 1 | 0 | 1 |
| R2 | 0 | 1 | 1 |
| R3 | 1 | -1 | 0 |
| R4 | 0 | 1 | 1 |
| R5 | 1 | 0 | 1 |
| R6 | 0 | 1 | 1 |

Figure 1.6: Illustration of the double-description based method for EFM calculation. The toy reaction network is reproduced from Figure 1.5. The stoichiometric matrix is converted into the irreversible form, through the separation of reaction 3 (R3) into forward and reverse components. The null-space of the irreversible stoichiometric matrix is calculated. The double-description method begins considering rows from the top of the null-space (R1, R2, ... , R6) until a row with a negative coefficient is encountered. Here R3f includes a negative coefficient in $N_3$. $N_3$ is combined with all columns with a positive coefficient in row R3f (here $N_1 + N_3$, $N_2 + N_3$), and is removed, (bottom middle). This combinatorial approach potentially leads to a rapid increase in the number of putative modes which must be considered. $N_1$ is removed during the recompilation of reversible reaction, leaving the three efms shown in Figure 1.5.

approach due to the ease with which various additional constraints can be added in this framework.

### MILP based approaches

As far as we are aware, de Figueiredo et al. [47] provide the earliest example in which the calculation of elementary modes is posed as an optimisation problem. In their approach, a binary variable $z_i$ is assigned to each reaction, $i$, such that $z_i = 1$ if reaction $i$ is active in the EFM, and zero otherwise. Each reaction is also associated with a non-negative flux $t_i$. They apply constraints

$$t_i \leq M z_i \quad \forall\, i = 1, ..., R \tag{1.2}$$

$$z_i \leq t_i \quad \forall\, i = 1, ..., R \tag{1.3}$$

to link $\vec{z}$, and $\vec{t}$, in order to ensure that no reaction can carry flux in an EFM unless it is included in that EFM (Equation 1.2), and that it must carry flux if it is included (Equation 1.3). $R$ is the total number of reactions in the system. $t_i$ can take any value less than $M$, where $M$ is some large scalar value (typically values of either 1,000 or 10,000 are used). Reversible reactions are decomposed, and constrained such that a reaction cannot carry flux in both directions in a single EFM, as

$$z_\alpha + z_\beta \leq 1 \tag{1.4}$$

where $z_\alpha$ and $z_\beta$ are the decomposed form of a single reversible reaction. They imposes the steady state constraint

$$\sum_{i=1}^{R} S_{c,i} \cdot t_i = 0 \quad \forall\, c \in I \tag{1.5}$$

for all metabolites $c$ in the internal set $I$. In order to avoid the trivial, zero-flux solution, they specify that at least one reaction must carry flux

$$\sum_{i=1}^{R} z_i \geq 1. \tag{1.6}$$

Equation 1.2 to Equation 1.6 define the steady state flux problem for a metabolic network $S$. To find EFMs, rather than all steady state solutions, they initially calculate the shortest EFM, by solving

$$\text{minimise} \sum_{i=i}^{R} z_i, \tag{1.7}$$

which returns the steady state solution involving the fewest number of reactions. This must be an EFM, as it cannot be decomposed into a smaller steady state solution.

Having now expressed the problem as one of optimisation, and found the shortest EFM, de Figueiredo et al. [47] extended their method to find the K-shortest EFMs by imposing additional constraints, preventing any returned solution from containing any previously found EFM

$$\sum_{i=1}^{R} z_i^k z_i \leq \left( \sum_{i=1}^{R} z_i^k \right) - 1 \qquad \forall\, k = 1, ..., K-1. \tag{1.8}$$

Where $z^k$ is a previously found EFM. This ensures that each found solution cannot be decomposed into smaller elementary modes, and therefore is itself a minimal solution (an EFM).

Although the enumeration of specifically short EFMs is of dubious biological relevance [47, 210, 66, 64], expressing the calculation of EFMs as an optimisation problem has allowed a suite of developments, due to the easy addition of constraints, and sequential output of EFMs, which allows subsets to be calculated. For example constraints can be easily applied such that only EFMs involving a particular reaction, $p$, are calculated [46, 162], simply by specifying the additional constriant

$$z_p = 1. \tag{1.9}$$

More sophisticated constraints have been applied to calculate only EFMs which are most likely to contribute to experimental flux, or transcriptomic data [162, 95, 176]. Further work in which metabolomic data, as well as thermodynamic constraints have been imposed upon the optimality problem, have indicated that between 50% and 90% of elementary modes are not thermodynamically feasible, depending on the organism, and methodology used [66, 64], allowing a reduction in the scope of the problem. However, given that organisms are likely to change the EFMs utilised depending on environmental conditions, these approaches may undermine many of the applications of EFMs for understanding the limits of the metabolic capacities of the organism . Where available, boolean regulatory models of organisms can be used to impose reaction dependence constraints. This has been shown to reduce the number of feasible EFMs by 99% [99], without compromising the EFM representation of the capabilities of the network under all conditions. However, it is possible that this approach is overly prescriptive, given uncertainty in the regulatory models themselves, and unfortunately such regulatory models are available for only a very few organisms.

Interestingly, FBA-type optimality constraints can also be applied to this formulation of the problem [104]. Although this shares the weakness of FBA, that some objective for the metabolic flux distribution is assumed, this can result in a more complete description of the optimal solution space than is possible through FVA, which loses all reaction dependency information.

In addition to drawing subsets which may be enriched for biologically utilised modes relative to the full set, it is also possible to attempt to calculate subsets which are able to approximate the behaviour of the full set, and thus to by

step the need for the enumeration of all modes. This has led to an interest in trying to enumerate a subset of EFMs which are diversely spread across the permissible steady state flux space. Kaleta et al. [100] used a genetic algorithm, in which constraints as to which reactions were permitted to carry flux, were imposed so as to preferentially calculate EFMs which were unlike those previously calculated.

We also developed an extension of the constraint-based de Figueiredo method [47] designed to give better coverage across the EFM space. This was implemented through a constraint additional to those used by de Figueiredo et al. [47], limiting the degree of similarity between a found EFM, and those previously returned,

$$\sum_{i=1}^{R} Z_i^m z_i \leq v \tag{1.10}$$

in which $v$ is the scaler number of reactions allowed in common with all previous found solutions, and $Z_i^m$ is the union of reactions in previously found EFMs, such that

$$Z_i^m = \begin{cases} 0, & \text{if } \sum_{p=1}^{K} z_i^p = 0, \\ 1, & \text{otherwise} \end{cases} \tag{1.11}$$

where $K$ is the number of EFMs previously found. By sequentially solving this linear model for $v = (0, ..., R)$, we find steady state solutions with as little reaction overlap as possible with the previously found EFMs, therefore comprising a diverse set of EFMs. This still guarantees that any returned solution must be an EFM, as it cannot be decomposed into other EFMs; if a solution of overlap $v$, were decomposable, then the smaller EFM must have overlap $\leq v$, and length strictly less that the current solution, $\sum_{i=1}^{R} z_i$, and yet not have been previously returned.

As is shown in Figure 1.7, this 'most diverse' approach did indeed lead to an improvement in the quality of small subsets of EFMs, as assessed by correlation between reaction participation in the subset, and the full set, calculated for a small model of *B. cenocepacia* J2315 metabolism [54]. Interestingly, the 'diverse' approach still performs significantly less well than a truly random sampling of the EFM population, presumably as a consequence of bias imposed by the minimisation of the number of participating reactions for a given permitted overlap (Equation 1.7), in order to guarantee that the returned solutions are EFMs.

Unexpectedly, as is shown in Figure 1.8a, in the model used, there was a strong correlation between the order in which EFMs were enumerated in the shortest approach, and our diverse approach, a consequence of the fact that many short EFMs also exhibit comparatively little overlap with previous solutions in comparison to long ones. This fact, in conjunction with the surprising increase in the average time required to find an EFM under the diverse constraint set (as shown

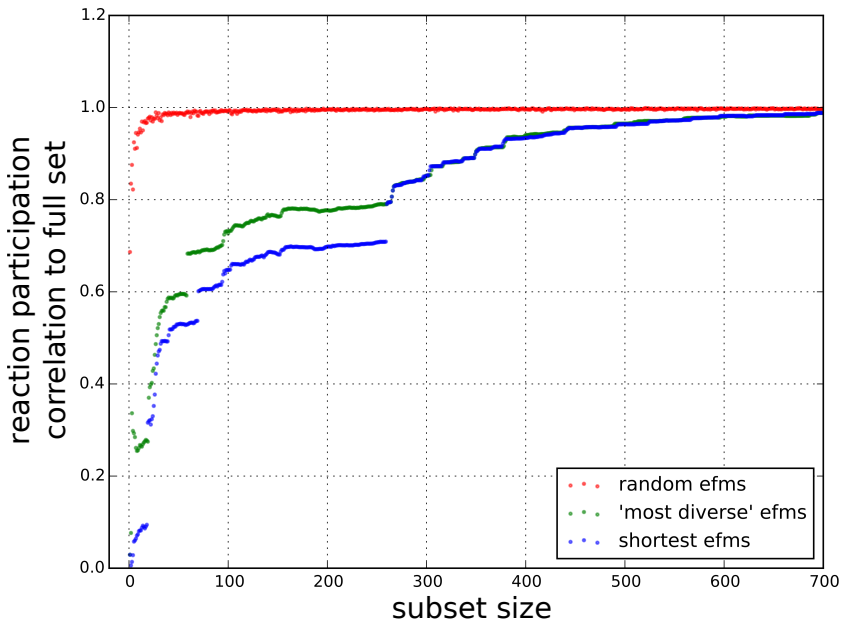



Figure 1.7: The 'most diverse' EFMs approach produces higher quality EFM subsets, than the shortest EFMs approach. Correlation in the fraction of EFMs each reaction participates in between the full set, and subsets generated by random sampling, the 'diverse' approach, and the shortest approach. The 'diverse' approach performs better than the shortest method for relatively small subsets, as there is greater correlation between reaction participation in the subset and full set, for subsets of less than approximately 260 EFMs. Both methods perform significantly less well than true random sampling from the population of EFMs.

in Figure 1.8b), meant that although our diverse approach performed comparatively well when EFM subsets of the same size are compared, it performs poorly relative to the shortest approach [47] when permitted an equal time for EFM calculation. Consequently, although this is an apparently obvious extension to the MILP family of approaches, it turns out not to be a practical one, in its most naive implementation.

While we were working to optimise this diverse approach, a dramatic improvement in constraint-based methods for EFM calculation was published [163]. Rather than MILP, Pey et al. realised that elementary modes could be equivalently calculated using a much faster, purely linear programming approach, without the need for the integer $z$ vector, whilst maintaining the benefits of the MILP framework. Although still slower than double-description based methods, this allows the rapid enumeration of relatively large subsets of elementary modes. In chapter 3 we use their TreeEFM tool to generate, to our knowledge, the largest set of elementary modes ever used for an analysis of plant metabolism.

## 1.5 Conclusion

In this chapter, we have introduced the concepts, and modelling frameworks relevant to genome scale models of metabolism. We have seen that although these approaches have been successfully applied, particularly in simple organisms such as bacteria, their use in plants, particularly outside of central carbon metabolism has been relatively rare. The aim of my PhD research has been to apply these methods to a genome scale model of *Arabidopsis thaliana* in order to gain insight into the mode and purpose of metabolic flux in this model organism, with particular regard to nutrient requirements. Here, we describe the further curation of a previously published constraint-based model, and its application to the study of sulfur starvation, and secondary metabolism. We then use a large set of EFMs to study the organisation of reactions into metabolic pathways, to investigate the relationship between internal metabolism, uptake fluxes, and nutrient use efficiency.

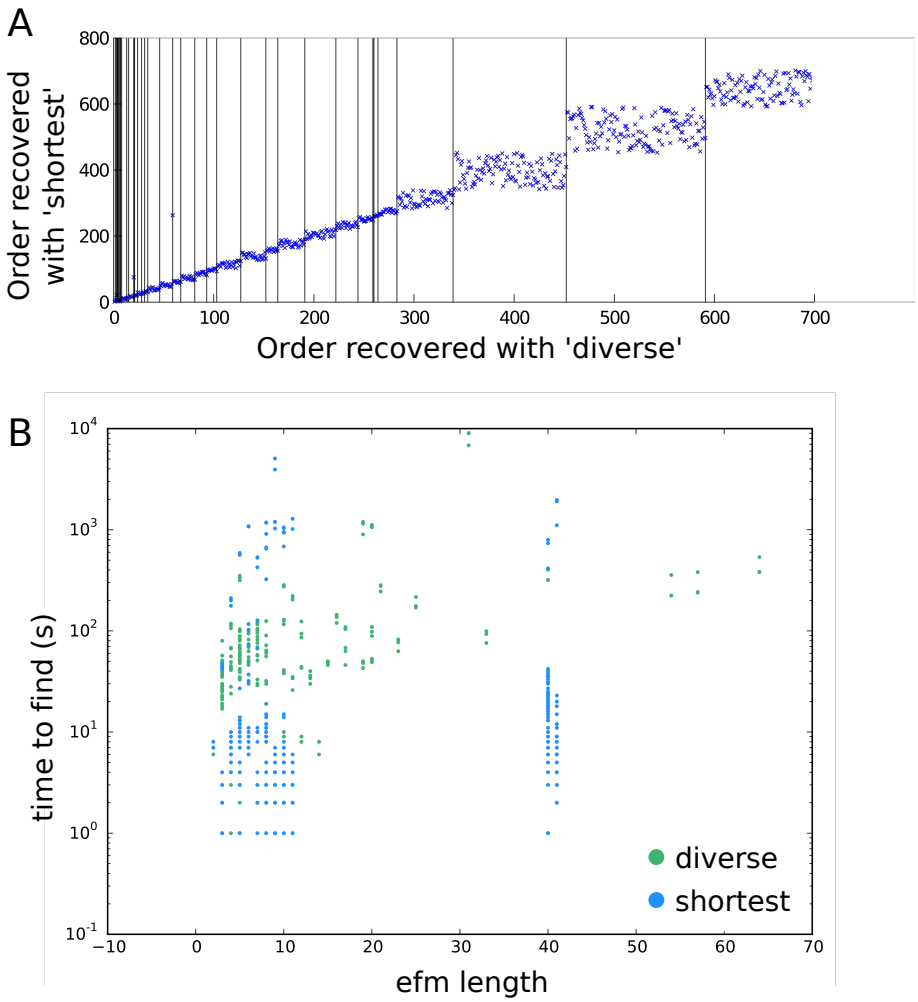



Figure 1.8: The 'diverse' approach performs poorly relative to the shortest per unit time. A, In the evaluated model, the order in which EFMs are returned are significantly related in the shortest, and 'most diverse' approaches. Vertical lines indicate points at which no more EFMs with the given overlap could be calculated, and the reaction overlap permitted was increased by 1. This is a consequence of shorter EFMs exhibiting less overlap with the calculated set on average than longer EFMs. B, the 'diverse' method takes longer to find each EFM than the shorter method. This is both because it tends to return longer EFMs, which are relatively slow to calculate under both methods, but also because the additional constraint results in all solutions being slower to find regardless of length.

# Chapter 2

# Model curation & flux balance analysis

Genome scale models of organisms consist of networks of metabolic reactions, often mapped to associated genes, and act as repositories of knowledge, similar to databases, but also can be used under various mathematical conventions to predict flux through internal reactions. A number of genome scale models of *Arabidopsis thaliana* have been published, and used for the analysis of central metabolism.

Here we further develop one of these published models, and use it to investigate genes involved in secondary metabolism and the metabolic response to sulfur starvation.

## 2.1 Previously published models

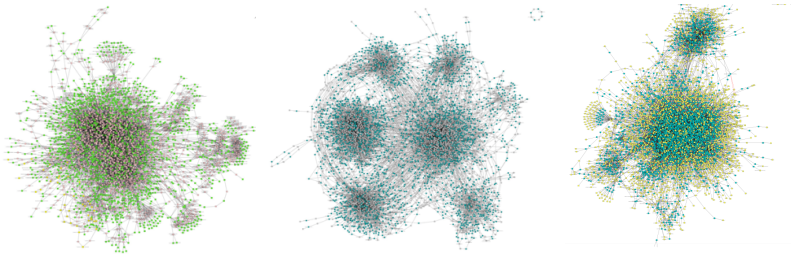
At the time that this work was carried out, several genome scale models of *Arabidopsis thaliana* metabolism had already been published. In order to determine which model to use for the study of sulfur starvation, we initially assessed the quality of these models. As our interest was in the effects of nutrient starvation, we focused particularly on the quality of the nutrient, and energy import-export requirement predicted by the models to be essential for growth. The oldest published model of Arabidopsis metabolism [169], focuses only on central metabolism of heterotrophically growing cell suspension cultures. This model was not considered to be compatible with our interests, due to its small scope, and is therefore not further discussed.

The AraGEM model [42] was the first compartmentalised genome scale model of Arabidopsis published. It consists of 1,748 metabolites and 1,567 reactions,

and is focused on primary metabolism. We assessed a slight modification of the AraGEM model which had previously been used to study glucosinolate metabolism [17], and was therefore considered likely to more fully incorporate sulfur reactions.

Using flux balance analysis, we saw no difference in the model prediction of the maximum biomass flux which could be produced under a 'full nutrient' regime, (in which import of Photon, Glucose, Maltose, and Sucrose were permitted), and a 'starvation' regime (in which they were not). This is likely to be because the AraGEM model is unable to synthesise all amino acids. and so is permitted to import Alanine, Aspartate, Glutamine, and Glutamate, which it catabolises under starvation conditions. As this model is not able to synthesise all biomass precursors from inorganic nutrients we considered it unsuitable for use by us in investigating plant nutrient requirements.

The Mintz-Oron model of Arabidopsis [145] was constructed using a novel, semi-automated approach from the KEGG and AraCyc databases. It is the largest of the models considered, and exhibits the most extensive gene-reaction mapping, and, (and as shown in Figure 2.1), compartmentalisation of reactions. However, the Mintz-Oron model predicts non-zero flux through the biomass reaction, even without access to any high energy substrates. This is equivalent to predicting that the plant should be able to grow in darkness, without access to any carbon source, and is caused by thermodynamic errors in permitted reaction directionality. It was therefore considered unlikely to be suitable for the study of nutritional requirements.



(a) AraGEM (b) Mintz-Oron (c) Cheung

Figure 2.1: The considered model reaction networks visualised using Cytoscape [199]. The superior compartmentalisation of the Mintz-Oron model can be seen as the separate, highly interconnected clusters of nodes.

The most recently published model at the time that this work was carried out was the Cheung model [33]. This is based on the AraCyc database, and has undergone extensive manual curation for reaction stoichiometry and reversibility. Table 2.1a shows that the Cheung model appears to reasonably predict inorganic nutrient requirements for growth, and correctly predict the capacity for growth using a variety of carbon sources. It also predicts sensible gas exchanges under autotrophic and heterotrophic metabolism, (Table 2.1b).

The only incorrect assessed nutrient requirement is that the model does not recognise that iron is essential for biomass production. Iron is not well incorporated into the model, as it is predominantly used as an enzyme cofactor, and the production of enzymes are not explicitly included in the model. The few reactions in the model which do include iron are either unable to carry flux, or are members of futile cycles, which are unlikely to occur in reality.

Since the input:output requirement predictions of the Cheung model were found to be markedly superior to other previously published models, we have used it as the basis of the rest of the work described here. Since this work was carried out, another Arabidopsis model has been published [5], however we have not assessed the quality of this model directly.

Table 2.1: Required environmental exchanges in the Cheung 2013 model. Uncommonly among published genome-scale models of Arabidopsis, the Cheung model correctly predicts most nutrient requirements for the production of biomass, predicts the requirement of an energy source, and is able to utilise a variety of energy substrates. It also correctly predicts the net direction of gas exchange under autotrophy, and heterotrophy.

| Energy Source | biomass? |
|---|---|
| Glucose | Yes |
| Sucrose | Yes |
| Starch | Yes |
| Sucrose | Yes |
| Photon | Yes |
| — | No |

| Unavailable nutrient | biomass? |
|---|---|
| $NO_3^-$ | No |
| $SO_4^{2-}$ | No |
| $PO_4^{3-}$ | No |
| $K^+$ | No |
| $Ca^{2+}$ | No |
| $Mg^{2+}$ | No |
| $Fe^{3+}$ | Yes |

(a) Energy and inorganic substrate requirements for biomass production

| Energy Source | $O_2$ | $CO_2$ |
|---|---|---|
| Photon | Export | Import |
| Sucrose | Import | Export |

(b) Gas exchange direction under autotrophy and heterotrophy.

## 2.2 Model validation and improvement

### 2.2.1 Gene knockout predictions

We have seen that the Cheung model makes simple, largely correct predictions as to the environments under which the plant is expected to be able to grow or

not. This is a common approach to validate the quality of bacterial models [55], however, the flexible, heterotrophic metabolisms of bacteria allow a much more rigorous assessment, due to the more varied environments in which bacteria can and cannot grow. This crude analysis therefore does not guarantee that the Cheung model is of high quality, and we wished to assess it in more detail, and potentially to develop it further. We therefore used additional approaches to assess the quality of the model.

A common approach to model validation, particularly in bacterial studies [51], but also in of other organisms [116], is to assess the quality of the growth / no growth model predictions in response to genetic perturbation, most commonly in response to single gene knockouts. We therefore compared the predictions of lethal single gene knockouts in the model to the database of Arabidopsis knockouts published by Lloyd & Meinke [128].

This dataset consists of information on the effect of 2,400 gene knockouts on growth. Of these we considered knockouts of the 270 genes annotated within the dataset annotated as being associated with 'metabolism'. This is because other essential processes, such as 'DNA and RNA synthesis', and 'chromosome dynamics' are beyond the scope of the model. Only 115 of these genes are believed to exist in a single copy in the Arabidopsis genome, based on the lack of sequence similarity to other proteins (BLASTP, e-30 cutoff). Non-unique genes were not considered, as there is no easy way to assess the contribution of the different gene copies to the overall gene function. Conversely we can reasonably assume that enzyme catalysed reactions cannot occur in the absence of single copy genes. Of the 115 unique genes, 62 could be manually mapped to reaction(s) included in the model. During this manual mapping, we also discarded from consideration gene products which are thought to function as part of an enzyme complex, but which are not essential for the (partial) functioning of this complex.

In Figure 2.2 we show the performance of the model for predicting genes which are experimentally essential, or inessential for growth. Knockouts were simulated *in silico* by constraining flux through reaction(s) catalysed by the gene product to 0, and assessing whether biomass could still be produced, such that flux solutions exist with non-zero flux through the biomass equation.

The left-most bar shows that the original Cheung 2013 [33] model performs perfectly in predicting the knock-out effect of genes, when simulating genes which are experimentally non-lethal. That is to say, knocking out these genes *in silico* also does not prevent biomass production. However, the right hand, pink bar shows that the quality of the predictions of lethal mutations is not so good, that is *in silico*, for about 75% of experimentally lethal genes, biomass can still be produced, even when reaction(s) which should be essential for growth are prevented.

Failing to correctly predict a lethal mutation is a consequence either of erroneous metabolic flexibility, or of failing to recognise that the production of a particular

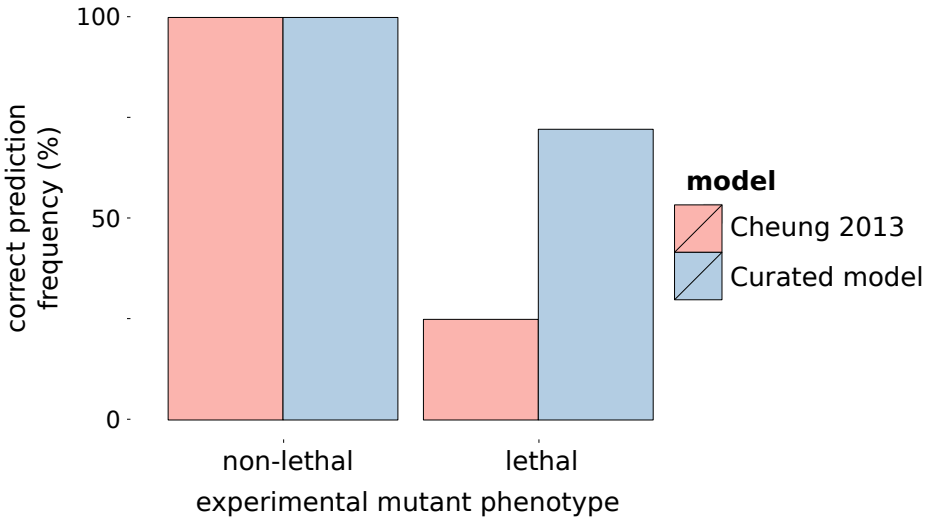



Figure 2.2: Accuracy of gene effect predicitons. 36 mapped genes knockouts were experimentally lethal, 26 genes were experimentally non-lethal. Model curation leads to improvement in the accuracy of prediction of lethal gene knock-outs.

metabolite, is in fact, essential for growth. Consequently, many of these errors were found to be corrected by addition of metabolites to the biomass reaction (see Table 2.8, in Methods, for details of the original biomass equation, and the metabolites added). The original biomass equation included in Cheung et al. is derived from chemical analysis of freeze dried cells [169, 244, 33], and it is therefore unsurprising that it did not include comparatively rare components. This highlights an interesting grey area in determining which metabolites are essential 'products' of metabolism, and which are just intermediates, produced in order to be converted into some essential product.

We are not aware of any rigorous solution to this problem, and the metabolites which were added to the biomass equation in order to improve gene lethality predictions are somewhat subjective. Anecdotally, we noticed that the addition of a metabolite to 'fix' an incorrect essential gene prediction often led to the introduction of errors in the prediction on non-lethal interactions, and it is likely that as bigger knockout datasets, and higher quality models become available, an iterative procedure to generate a parsimonious biomass equation can be adopted.

Although the metabolites added to the biomass equation must be producible by the plants metabolism, we had no experimental data about the quantity which is required for growth, and therefore added them at a nominal, small coefficient (1e-6) to the biomass equation. We therefore do not expect these additions to make much difference to internal flux distribution. What was informative however was that a large fraction of these metabolites could not be produced

by the original Cheung model. This is a consequence of missing reactions; through manual curation, and after examination of the literature, we added 14 chemical reactions, and 2 inter-compartmental transporter reactions to allow their production.

This process also highlighted the existence of some phenotypes which cannot easily be addressed using a steady state constraint-based approach. For example 'UPP' is an experimentally essential gene, which catalyses a reaction in the in the Pyrimidine salvage pathway, however it was not predicted to be essential *in silico*. In fact, the associated reaction did not carry any flux, even when unconstrained in the optimal, maximum biomass flux solution found. It is not clear that standard flux balance analysis will ever be able to identify salvage pathway genes as being essential, because rather than having to salvage chemicals which have previously been produced above the currently required level, the simulated solution will simply produce only the required amount of the chemical, saving the energy expended through recycling metabolites.

Even after further curation (blue bars in Figure 2.2), the apparent quality of predictions for non-lethal mutants is markedly greater than the quality for essential genes. However, this a consequence of many reactions not carrying flux in the optimal solution instance found, and so not affecting biomass production when they are constrained to carry zero flux. This is not in itself necessarily a sign of the model being 'incorrect', as these reactions could exist so as to provide metabolic flexibility either to genetic, or environmental perturbation [157], however many of these reactions *cannot* carry flux, even when the objective function is to maximise flux through that reaction. This could potentially be a consequence of the environmental constraints, but is more likely to be caused by an incomplete reaction network. In the next section, we describe further curation of the model to address the presence of these 'blocked' reactions.

### 2.2.2 Blocked reactions

It is well known that the 'optimal' flux distribution found by flux balance analysis often involves only a small fraction of the reactions included in the model. This is predominantly ascribed to the need for metabolic networks to have evolved so as to be robust to genetic and environmental perturbations, and so to have a somewhat redundant set of reaction systems which are not necessarily used in a particular, studied, environment, but which may carry flux under other conditions [157].

Consistent with this, we found that in the flux solution returned with the objective of maximising flux through the biomass reaction, only 573 reactions of the 2,799 reactions in the model have non-zero flux. However, we also found that only 445 of the reactions which don't carry flux in the optimal solution can carry flux, when maximisation of flux through them was designated as the objective function, meaning that 1,781 of the 2,799 reactions in the model cannot carry

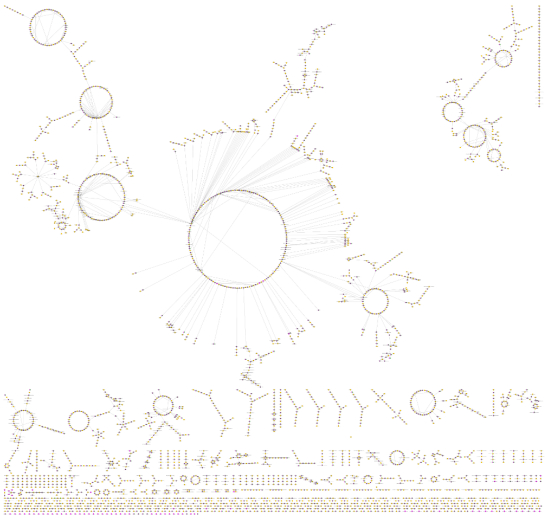
flux at steady state. We also found that 1,874 of the 2,625 metabolites in the model are only involved in these reactions which cannot carry flux. It is possible that with a different permitted set of external metabolites, these figures may change somewhat, however, given that the used external set reflects the nutrients commonly available to a plant, we consider it unlikely that this is the main cause of this result. Instead the likely cause is the inability to produce substrates for the blocked reactions, or to consume their products, itself caused by an incomplete reaction network.

Not all blocked reactions are themselves problematic, most can be expected to be up, or downstream of some 'root' problem reaction which cannot carry flux, for example due to an unconsumed, non-exported product, and which then leads to knock-on effects. To identify the root, 'causal' reactions, we visualised the blocked reaction-metabolite sets as clusters, by grouping contiguous sections of blocked reactions together. Blocked metabolite-reaction clusters are shown in Figure 2.3. Reactions at the edge of a cluster are candidate 'causes' of the blockage of the whole cluster, whereas reactions in the middle of a cluster are likely blocked only as a consequence of reactions at the edge. This approach reduced the number of problems which had to be examined from 1,781 blocked reactions, to 539 blocked reaction clusters, and allowed us to focus manual curation efforts on the likely roots of the problems. Through manual curation of the blocked reactions, in order to allow them to carry flux, we added 174 reactions, removed 1 erroneous reaction, and added a further 32 metabolites to the biomass equation (shown in Table 2.8 in Methods). This led to a reduction from 539 blocked reaction clusters to 271.

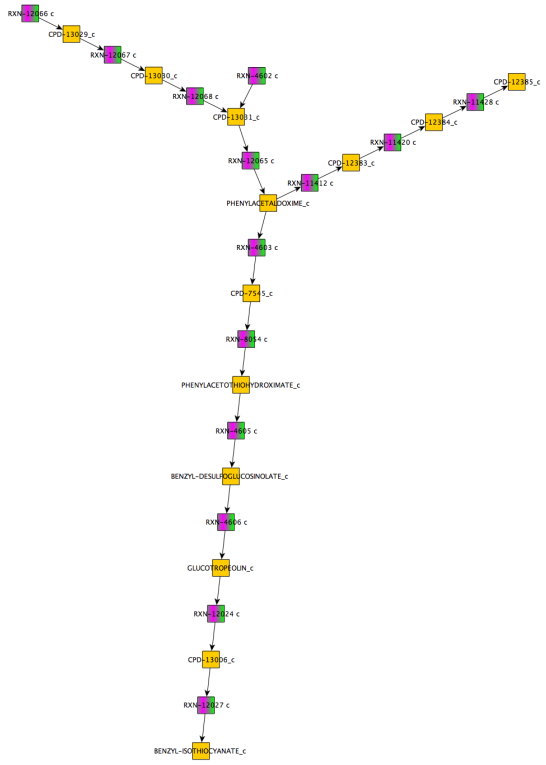
### 2.2.3 Flux prediction comparison

Having spent a significant amount of time in improving the qualitative predictions of the model, we wished to check the impact of the changes on quantitative predictions. We assessed the differences between the original model published by Cheung et al. [33], and the curated model, in predicting internal fluxes. These were estimated by flux variability analysis (FVA)[132]. We compared the FVA predictions of both models to experimental fluxes estimated through metabolic flux analysis using tracer experiments, as is commonly done in the literature. The Cheung model had previously been shown to exhibit a good fit to these kinds of data [33], however, we wished to confirm that the changes made had not compromised these predictions, and also compare the models to a wider range of experimental datasets.

The non-growth associated ATP maintenance cost of a cell is essential for accurate flux predictions. However it is difficult to directly measure, and is often fitted to the measured uptake of energy providing substrates, and growth rate [169, 82]. Here, inspired by this approach, we varied the ATP maintenance cost, and, in Figure 2.4, plotted the correlation between FVA, and MFA estimated flux across all measured reactions for each ATPase flux value. For each dataset



(a) All blocked clusters.




(b) An enlarged example of a blocked cluster. Yellow boxes are metabolites, purple and green boxes are reactions.

Figure 2.3: Visualising clusters of blocked reactions and metabolites allows manual curation to focus on the causes of the blockage. From the Cheung model; clusters of reactions which cannot carry flux, as flux variability analysis (FVA) [132] predicts that maximum and minimum flux equals zero, and metabolites which are only involved in blocked reactions. Directed edges indicate the metabolites produced and consumed in blocked reactions. Although all shown nodes are blocked, the cause of the blockage is normally a reaction, or metabolite on the edge of each cluster.

(facet in Figure 2.4), we saw that the results for the Cheung and modified models are largely superimposed, indicating that, as expected, the modifications made very little relative difference to the predictions of flux through central metabolism in the original and modified models. It is encouraging that across all datasets, the modified model either performed identically, or with slightly improved correlation to the MFA data relative to the Cheung model, as is shown by a greater maximal coefficient of determination for the modified model in each facet. Another encouraging difference is that in the Cheung MFA data facet, the best model performance (correlation), is achieved at slightly lower ATP maintenance flux in the modified model, as it was previously slightly overestimated [33].

Interestingly, Figure 2.4 shows that the quality of flux predictions of both models varies substantially between datasets. In five of the studied metabolic flux analysis datasets (Cheung 2013 [33], Masakapalli 2010 [139], Williams 2008 [243], elevated and standard oxygen, Williams 2010 [244]), we see reasonable correlation between the flux balance analysis and metabolic flux balance analysis results. All of these datasets were generated using heterotrophic Arabidopsis cell-suspension cultures, and focus on flux through reactions in central metabolism. In comparison, model predictions are less similar to datasets generated using illuminated Arabidopsis rosettes (Szecowka 2013, [215]), or focused on reactions involved in the production of cell wall precursors (Chen 2013, [29]).

The relatively low correlation between FBA and the MFA data generated by Chen et al. [29] is likely to reflect the relative accuracy with which different parts of metabolism are represented in the model. It suggests that reactions not involved in central carbon metabolism are not as well represented as those which are, and also that the requirements for cell wall precursors are not as well reflected in the biomass equation as other components. The different experimental focus of the Chen 2013 dataset [29] perhaps explains why the largest improvement between the original, and modified models across datasets is seen in it. The relatively small changes we have made to the biomass equation, and reaction network are unlikely to affect the (comparatively large) fluxes through central metabolism significantly in comparison to the (comparatively small) fluxes in other parts of the network.

There is also relatively low correlation between FBA, and MFA in the Szecowka dataset [215]. It is important to remember that the model is essentially of a single cell. There are therefore discrepancies between the reaction structure of the model and the metabolic conversions of a complete Arabidopsis rosette, for example transportation between tissues is not explicitly incorporated in the model. However, this discrepancy is also related to difficulties in the assumptions of the flux balance analysis method. Although simple objective functions can be used to generate reasonable estimates for flux in bacteria, and cell-suspension clusters, as discussed in chapter 1, it is not clear that a simple linear optimisation problem can be mapped onto the 'true' objective of more complex organisms. Even bacterial fluxes appear to exist within some objective tradeoff space, and

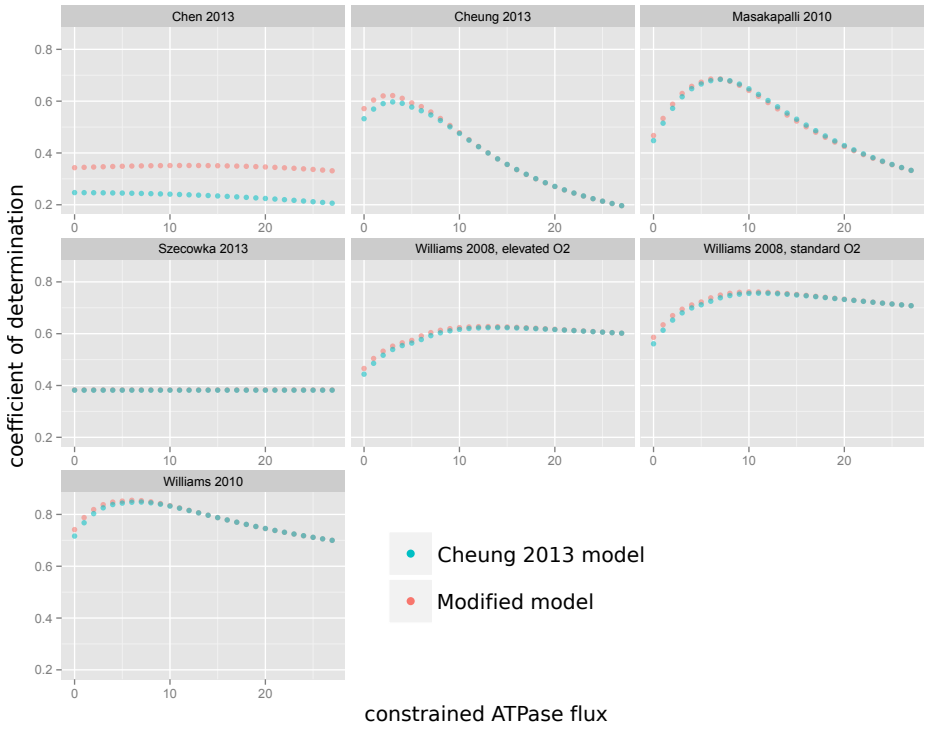



Figure 2.4: Flux predictions for reactions in central carbon metabolism using the curated model are largely unchanged from the original Cheung model. The coefficient of determination ($r^2$) between metabolic flux analysis flux, and flux balance analysis flux in the Cheung 2013, and modified models we calculated over varied maintenance ATPase flux constraints. Flux balance analysis values were taken as the mean of the upper and lower reaction flux bounds found by FVA, as suggested in [205]. All fluxes were normalised by uptake flux. There is some improvement in agreement between the the curated model and the MFA estimated fluxes compared to the original model, but it is not clear that it is a significant improvement, given experimental uncertainty, and particularly, the crudity of the FVA approach. There is some variability in the optimal maintenance flux which fits the data best between experiments, as indicated by the greatest coefficient of determination being at different constrained ATPase flux values in the different facets. This is possibly a consequence of different metabolic requirements under different experimental conditions.

it is likely that more complex organisms with multiple cell types will be even more complex, and in particular that different tissues may be best described using different objectives. As such measured fluxes, averaged over the whole organism, or even at the organ scale, are potentially difficult to recover using flux balance analysis based methods.

In the particular case of the Szecowka dataset [215], mature rosettes are unlikely to build much additional biomass, and instead predominantly produce intermediate metabolites for export to the rest of the plant. It is likely that altering the biomass equation to better reflect the function of photosynthesising leaves would improve the quality of the fit, although it is unlikely that it will be as good as in the other datasets.

Figure 2.4 confirms that the quality of the flux predictions of the Arabidopsis model has not been compromised by the modifications we have made in order to improve other, (qualitative), predictions. We therefore now explore its suitability, and the suitability of the FBA method for less introspective applications.

## 2.3 Genes which affect glucosinolates

### 2.3.1 Introduction

One use of genome scale models in bacteria is for the rational design of genetic engineering strategies, in order to increase the production of metabolites of interest [156]. In principle, the same approach can be used for the design of rational genetic intervention strategies in plants. However, the majority of commercially interesting metabolites are products of secondary metabolism, which in comparison to primary metabolism is relatively poorly understood, and genome-scale models of Arabidopsis have historically focused on central carbon metabolism [41]. In order to assess the suitability of applying FBA, and the Arabidopsis model to secondary metabolism, we investigated the quality of predictions of the genes which affect the production of glucosinolates.

We focused on glucosinolates, due to 1.) their status as near 'model' secondary metabolites [206], the biosynthesis of which is comparatively well understood. This is indicated by the fact that a genome scale model has previously been used for analysis of their production [17]. 2.) Glucosinolates integrate carbon, nitrogen, and sulfur metabolism, and therefore their production is potentially influenced by a large number of metabolically distant reactions, which could be non-intuitive, but accessible through FBA analysis. 3.) Modification of glucosinolate profiles is of interest due to their role in pest resistance [232], and potential nutritional benefits [221].

### 2.3.2 Comparison of FBA predictions to genes known, and expected to affect glucosinolate production

Glucosinolates are produced from amino acids, and divided into three classes according to their amino acid precursor. Aliphatic glucosinolates, are derived from methionine, indolic glucosinolates are derived from tryptophan, and benzenic glucosinolates are derived from phenylalanine or tyrosine. We first ensured that the model was able to produce a wide variety of indolic, aliphatic, and benzenic glucosinolates (see Table 2.14 in Appendix),

In order to initially assess the quality of glucosinolate related predictions derived from FBA of the Arabidopsis model, and therefore validate this approach, we compared model predictions regarding genes involved in glucosinolate production, to genes which are known or expected to affect their production. Specifically, we compared those genes, which when knocked out *in silico*, are predicted to alter the capacity for glucosinolate production, without significantly compromising biomass production (see Methods), to experimentally identified genes [28] which have been shown to affect glucosinolate profiles, and to genes which are expected to affect glucosinolate production due to either biochemical, genetic, or homology based evidence, or through expert expectation (these genes are listed in Table 2.13 in Appendix, reproduced from [28]). This comparison set consists of 16 genes which are known to affect glucosinolate metabolism, and 152 genes which are expected to.

Although it is expected that genes outside of this table also affect glucosinolate metabolism, and it is ultimately these that we wish to identify, we considered this set as a 'gold standard' to initially assess the quality of predictions being made. This is similar to the approach taken by Chan et al. [28], who tuned the threshold criteria used to identify candidate genes in their Genome Wide Association Study (GWAS) to achieve the best performance in recovering the members of this set.

The results of this analysis are shown in the top row of Table 2.2. This shows that agreement between the predictions, and expected genes is quite high, specifically, we see that both precision, and recall to the expected gene set is greater for FBA derived predictions than was achieved by the GWAS approach [28].

The motivation for requiring that knockouts still be able to produce biomass, is that many knockouts which prevent the production of glucosinolates are in fact much more general, and prevent flux through much of the reaction network. For example it is not clear that knocking out a reaction which prevents uptake of the sole energy source available, thus preventing growth, is best considered as a mutant which affects glucosinolate production. By including a biomass requirement, we hoped to prevent the identification of these kinds of reactions, which could be considered false positives. However, comparison of the top two rows of Table 2.2 shows that not only is recall (the fraction of true positives identified) greater without the biomass requirement, but so is precision (the

Table 2.2: Flux balance analysis performs better at re-identifying expected glucosinolate genes than GWAS. 'Reaction KO'; reactions are individually prevented from carrying flux, genes associated with reactions which affect glucosinolate production are reported. 'Gene KO'; all reactions associated with each gene are prevented from carrying flux. '& biomass'; knock-out must still be capable of producing similar biomass flux to the 'wt' model. 'no biomass'; there is no requirement that the mutant must be able to produce biomass (see Methods). 'Correct predictions'; the size of the intersection between genes predicted by the FBA approach, and the 168 genes expected to be involved, based on homology, biochemical or genetic evidence, and biologically informed guesswork (listed in Table 2.13). Precision is the number of correct predictions divided by total predictions. Recall is the number of correct predictions divided by the number of expected genes. It can be seen that all FBA approaches used have greater precision and recall than the GWAS experiment carried out in [28].

| | 'correct' predictions | number of predictions | Precision (%) | Recall (%) |
|---|---|---|---|---|
| reaction KO, & biomass | 29 | 220 | 13.18 | 17.26 |
| reaction KO, no biomass | 56 | 413 | 13.55 | 33.33 |
| gene KO, & biomass | 23 | 315 | 7.30 | 13.69 |
| gene KO, no biomass | 74 | 803 | 9.21 | 44.04 |
| GWAS, 2007 | 12 | 1,056 | 1.13 | 7.14 |
| GWAS, 2008 | 11 | 893 | 1.23 | 6.54 |

number of true positives divided by predicted positives).

This is partially a consequence of the comparison between the simulation of total knockouts, to the expected gene set, which is based on the expectation of altered function, rather than the total knockout of the gene. It is also partially a consequence of the approach to mapping between reactions and genes; we considered that each gene associated with a reaction was essential for its function which is not necessarily true. For example the model considers knockout of each of the APS reductase isoenzymes to be lethal, although in reality they can partially compensate for the loss of each other. Consequently the expected gene table (Table 2.13), includes gene products as affecting the glucosinolate phenotype, which catalyse reactions which if knocked out completely would prevent the formation of biomass. We could exclude predicted reactions for which the gene:reaction mappings which are not 1:1 in the same way we previously did when comparing to the lethal knockout database in subsection 2.2.1, however this was found to severely compromise FBA recall, so as to offset any potential benefit.

Upon manual examination of the reactions which were not correctly predicted, we found that several false negative predictions, (i.e. genes which effect glucosinolate production, but were not identified by flux balance analysis), were caused by the existence of obvious parallel reactions in other subcellular compartments. To fix this error, rather than applying 'reaction knockout', we tried applying a 'gene knockout', in which all reactions associated with a given gene

product were simultaneously prevented from carrying flux. However although this did address this problem, it introduced some others. It is not likely to be true that each gene is essential for each reaction it is associated with, and so as is shown in Table 2.2, we consistently end up with more false positive predictions in the gene knockouts, as shown by the lower precision achieved. Although recall was higher in 'gene' rather than 'reaction' knockout when biomass was not also required, 'Gene knockout, & biomass' performs poorly, because genes in the expected gene table (Table 2.13), prevent the formation of biomass when knocked out, and so are not returned under these requirements.

Although it seems that dropping the biomass requirement does improve performance, it is not clear that either of the reaction / gene approaches used unequivocally performs best, instead they exhibit performance trade-offs. It is also important to remember that we are only making a comparison to an expected set, not the true set, and it is possible that this expected set itself includes some false positives, as well as the presumed false negatives.

To examine whether the prediction inaccuracies could be improved through further model curation, we looked in more detail at the identities of the expected genes which FBA failed to predict. We saw that a large number of the genes which are expected to affect the glucosinolate phenotype, were not identified by the flux balance analysis due to not being included in the databases used to map between gene identities and reactions (see Table 2.3). To find the cause for this, we examined the identities of the genes. Table 2.4 shows the genes which have previously been experimentally confirmed as involved, and which are fairly representative of the full (expected) set. This indicates that the majority of expected genes missing from the databases are beyond the scope of the reaction network model. We are interested in mapping genes to catalysed metabolic reactions, and therefore the databases used do not include transcription factors, and other genes associated with regulation. Regulatory processes are not incorporated into the model, and so could not be recovered by flux balance analysis, even if the missing genes were in the databases. Only six genes are associated with reactions which are not incorporated into the model.

Forty expected genes are included in the model, but not recovered by FBA using the 'gene KO, no biomass approach' (gene identities are given in Table 2.16 in Appendix). Of these, we see that 18 genes are associated with the breakdown of glucosinolates or their precursors. FBA cannot easily recover genes which are involved in metabolite breakdown or recycling, as typically these processes are not used in the optimal predicted flux distribution.

In at least one case, an expected gene was not recovered because the particular glucosinolates it is involved in producing were not exportable in the model. AOP2 is involved in the production of 2-propenyl-glucosinolate, and 3-butenyl-glucosinolate, however these glucosinolates could not be produced from inorganic substrates in the model, because it requires the production of an unconsumed methane-sulfonate side product. Although this reaction can be included in the reaction network structure of the model, it is so proximal to the pro-

Table 2.3: A large number of the expected genes were not recovered due to the absence of gene:reaction mapping information. 'Expected genes' is the number genes which are expected to cause to a glucosinolate phenotype (see Table 2.13 in Appendix). The databases used for mapping between genes and associated reactions were TAIR, and Biocyc, which only include 120 of the 168 expected genes. Six genes which are associated with reactions in these databases could not be mapped to the model due to reactions being absent from the model.

| | number of unique genes |
|---|---|
| expected genes | 168 |
| of these, genes associated with reactions in databases | 120 |
| of these, reactions in model | 114 |

duction of these glucosinolates that it is unlikely to lead to widespread errors. Consequently we conclude that modifications to the reaction network structure itself are unlikely to result in significant performance improvements.

Aside from potential errors/incompleteness of the modelled reaction network, it is possible that the other expected genes, which were not recovered by FBA could be due to falsely included expected genes, but this seems relatively unlikely. The expected genes are largely based on reasonable evidence, or very obvious biological intuition, essentially through informal metabolic flux analysis. For example many genes involved in methionine biosynthesis are expected, presumably because they are expected to affect the production of methionine, a glucosinolate precursor.

A more likely source of failure to recover expected genes is due the mapping of genes to reactions. This could be due to the incorrect, or simply incomplete mapping of genes to reactions. Although these genes are mapped to at least one reaction in the model, there is no guarantee that this means that they are correctly mapped, or that they are mapped to all of the reactions that they should be, and particularly the reaction which actually causes the phenotype effect. Additionally, as previously discussed, although the mapping between genes and reactions is not necessarily 1:1, we employed an extremely simple logic to relate them, in which all genes associated with a reaction were assumed to be essential for its function. Although this is obviously a source of error, a more sophisticated mapping approach requires extensive manual oversight, and was not considered a good use of time, as this simplification can only cause false positives, not false negatives when the production of biomass is not also required.

We have seen that FBA approaches appear to perform relatively well in recovering genes which are expected to affect glucosinolate metabolism. This agreement between FBA predicted, and biologically expected genes does somewhat validate the quality of the model. However, whilst this is encouraging, the relatively high quality of flux balance analysis based predictions in comparison

Table 2.4: Genes which are known to affect glucosinolate phenotypes are recovered well by FBA in comparison to GWAS, but also demonstrate some shortcomings of the approach. AGI shows the Arabidopsis Genome Initiative identifier for each gene, Evidence indicates the type of experimental evidence for its inclusion as a gene which affects glucosinolate phenotypes, Recovered by GWAS indicates whether this gene was recovered by Chan et al. [28], * indicates that it was recovered in one year only, ** indicates that it was recovered in both GWAS experiments. We see that a greater proportion of the genes which are already known to affect glucosinolate production can be recovered through FBA analysis than through GWAS. Of the genes which were not recovered by FBA, the majority are not recovered due to their absence from the genome scale model, most commonly because they are Transcription factors, which are not incorporated.

| AGI | Gene name | Pathway | Evidence | Recovered by GWAS | Reason not recovered |
|---|---|---|---|---|---|
| *Recovered by FBA* | | | | | |
| AT1G62540 | GSOX2 | Aliphatic Glucosinolate | Biochem | | |
| AT5G23010 | MAM1 | Aliphatic Glucosinolate | Biochem | ** | |
| AT1G12140 | GSOX5 | Aliphatic Glucosinolate | Biochem | | |
| AT1G62570 | GSOX4 | Aliphatic Glucosinolate | Biochem | * | |
| AT4G03050 | AOP3 | Aliphatic Glucosinolate | Biochem | ** | |
| AT5G57220 | CYP81F2 | Indolic Glucosinolate | Biochem | | |
| AT1G65860 | GSOX1 | Aliphatic Glucosinolate | Biochem | | |
| AT1G62560 | GSOX3 | Aliphatic Glucosinolate | Biochem | | |
| AT1G24100 | UGT74B1 | Aliphatic Glucosinolate | Biochem | | |
| *Not recovered by FBA* | | | | | |
| AT2G25450 | GS-OH | Aliphatic Glucosinolate | Genetic | | Associated reaction is not in the model |
| AT2G31790 | UGT74C1 | Aliphatic Glucosinolate | Biochem | | Gene id is not in database |
| AT5G07690 | MYB29 | Aliphatic Glucosinolate | Genetic | | Transcription factor |
| AT5G07700 | MYB76 | Aliphatic Glucosinolate | Genetic | | Transcription factor |
| AT5G60890 | ATR1/MYB34 | Indolic Glucosinolate | Genetic | | Transcription factor |
| AT5G61420 | MYB28 | Aliphatic Glucosinolate | Genetic | | Transcription factor |
| AT4G03060 | AOP2 | Aliphatic Glucosinolate | Biochem | ** | Produced glucosinolate cannot be exported in wild-type model |

to GWAS is perhaps unsurprising. The genome scale model is partially built through genetic studies as to which genes affect what phenotypes, and the reaction network structure it incorporates is at least partially the underlying data behind the biological intuition which identified many of the expected genes in Table 2.13. The high agreement between these two approaches must be partially considered a consequence of this somewhat incestuous methodology, and does not necessarily reflect the (potentially unexpected) genes which are truly important in controlling glucosinolate phenotypes.

Whilst the agreement with known and expected genes is encouraging, what is more interesting is to see whether the model is able to make correct predictions that are more obscure than those which are possible through intuition alone. This non-obviousness could be caused by metabolic distance (many of the expected genes are those which catalyse the steps immediately preceding glucosinolate formation), or conceptually, for example it is trivial to expect that sulfur assimilation genes might affect production of glucosinolates, the well known sulfur containing metabolites. Although non-obvious predictions are made (those FBA predictions which are not in Table 2.13 ), it does not automatically follow that these are of the same apparent quality as the obvious predictions we have already discussed. In the next section we therefore compare FBA predictions to GWAS, a somewhat unbiased method for identifying genes involved in production of glucosinolate profiles.

### 2.3.3 Comparison of FBA to GWAS

In genome wide association studies (GWAS), natural variation in the genetics of a population is used to assess the contribution of genetic loci to assessed phenotypes. This is potentially a powerful tool for the investigation of genes involved in a given phenotype, which is unbiased by previous knowledge. However, the underlying population structure, and linkage disequilibrium between causative, and unrelated alleles can make it a potentially quite error prone process [246].

We compared FBA predictions to the results of an unbiased GWAS into the loci which affect glucosinolate production [28]. This was done primarily in order to validate the FBA approach taken, by assessing whether non-obvious predictions truly affected glucosinolate production, but also because both GWAS and FBA methods are expected to be fairly error prone, with many false positive identifications. As GWAS and FBA are independent methods, relying on different assumptions, we hoped that by using both methods in conjunction we could to prioritise genes for further, targeted experiments, or at least generate hypotheses as to the processes by which genes identified by GWAS affected glucosinolate production.

Consistent with the idea that both methods result in large numbers of false positives, Table 2.5 shows that there is relatively little agreement in the genes

identified by GWAS, and FBA. However, there is some mutual consistency, as across all FBA modelling approaches used, the number of genes predicted by both methods is greater than the statistically expected overlap, which would be expected by chance if FBA and GWAS made totally independent predictions. This lack of complete independence is presumably because both methods are to some extend independently identifying the true genes which affect glucosinolate production, and somewhat validates both the FBA approach, and the particular GWAS study used in this comparison.

Table 2.5: There is more agreement between GWAS and FBA approaches than expected by chance. Statistically expected is the expected number of genes returned by both methods assuming random, independent gene picking; calculated as $\frac{p_{gwas}}{a_{gwas}} \cdot \frac{p_{fba}}{a_{fba}} \cdot g_{both}$ where $p$ is the number of predicted genes, $a$ is the number of genes considered in FBA (4,037), or GWAS (31,505), and $g_{both}$ is the number of genes in both the model, and GWAS study (2,609), Biologically unexpected is the number of genes predicted by both the GWAS, and FBA approaches which are not present in the biologically expected table (Table 2.13). The overlap between predicted genes is greater than would be expected by chance if the predictions were completely independent. This suggests that although the two methods are independent from each other in how they work, they do contain some mutual information, presumably because they both identify some true causal genes, validating both methods. However, in each case, the number of biologically unexpected genes is similar to the statistically expected overlap between the two methods. This suggests that the correct, mutual identifications concern genes which are already intuitive, and that therefore FBA modelling adds little information. The number of non-obvious predictions is similar to the number that is expected by chance; it is therefore not clear whether or not the non-obvious predictions made by both methods are actually more reliable than those predicted by only one, or are a statistically inevitable consequence of the number of predictions made by two independent error prone methods.

| FBA approach | Genes predicted by FBA | Genes predicted by GWAS | Genes predicted by both | Statistically expected | Biologically unexpected |
|---|---|---|---|---|---|
| rxn KO and biomass | 220 | 1646 | 15 | 7 | 7 |
| rxn KO no biomass | 413 | 1646 | 28 | 14 | 15 |
| gene KO and biomass | 315 | 1646 | 17 | 11 | 11 |
| gene KO no biomass | 803 | 1646 | 35 | 27 | 21 |

We saw that in this comparison, the simulated knockout of individual reactions, rather than gene knockouts apparently leads to greater accuracy, as assessed by the number of genes predicted by both FBA and GWAS, relative to the statistically expected overlap in Table 2.5. However, the reaction based knockout approach has a higher number of false negatives, as shown by the reduction in the number of genes predicted by both FBA and GWAS. The reduced accuracy of gene knockout approaches is presumably because of erroneous gene:reaction mapping, and the overly robust assumption that a gene is essential for the function of all associated reactions, as previously discussed, leading to many more genes being predicted to be involved in glucosinolate metabolism. We again conclude that none of the described FBA approaches perform unequivocally

'better' than the others.

It appears that the majority of the genes which are mutually predicted by GWAS and FBA approaches beyond the statistically expected number are due to intuitive, biologically expected genes. This can be seen as the number of biologically unexpected genes (the genes predicted by both FBA and GWAS, but which are not in Table 2.13) is broadly similar to the statistically expected number of mutually returned genes, based on the number of genes identified by each approach.

This could be a consequence of the relative scarcity of true causal genes; for example, in a conceptual scenario in which truly only 35 genes affect glucosinolate prediction, the mutual use of GWAS and 'gene KO no biomass' FBA approaches would have performed perfectly in eliminating the false positive predictions made by either method alone, and the similarity of 21 biologically unexpected genes to 27 statistically expected genes would be coincidental. However, this seems unlikely to be the case. Instead it is likely that the biologically obvious predictions of the FBA modelling are of higher quality than the non-obvious ones, to the extent that the non-obvious ones bear no more resemblance to the GWAS study (and by implication therefore biological reality) than random sampling. It is not clear therefore that the addition of FBA to a GWAS study adds much more than biological intuition to the credibility of, or mechanistic explanation for any GWAS result, and it is certainly not clear that the genes which are mutually predicted are any more likely to be correct than the genes which were only identified by GWAS.

We investigated in further detail why there is so little agreement between the FBA and GWAS approaches as to which genes are important, as depending upon the underlying cause we may have been able to refine and improve our approach.

We note that it is of course possible that GWAS is the inaccurate method, given its relatively poor performance in recovering expected genes. However, this is likely a consequence of its unbiased basis, which does not account for previous knowledge of glucosinolate production. In comparison, as discussed previously, FBA can be expected to perform relatively well in predicting expected genes, as the model structure is designed so as to incorporate this kind of biological knowledge, and relatively poorly in predicting genes which are non-obvious. Furthermore, GWAS is a well established method for the identification of novel trait loci. In comparison FBA, particularly in plants, has not been widely applied to this kind of analysis.

The surprising lack of agreement is partly due to differences in genes covered by the curated Arabidopsis model and the SNPs used for the GWAS. Of the 4,037 genes in the model, and 31,505 genes in the GWAS study, only 2,609 genes are present in both. Consequently, of the 1,646 genes predicted to cause a glucosinolate phenotype in the GWAS study, only 163 are included in the model at all. It is also important to remember that although these genes are in the model,

this does not mean that they are completely, or accurately included.

The difference in gene coverage goes some way to explaining the difference, but there is still little agreement as to the important genes, even among the ones which are included in both approaches. We speculate that this difference is primarily because the different methods are primarily able to identify different types of genes. FBA analysis can only identify enzymes, as regulatory genes are not included in the model. Conversely, examination of the most common gene ontology annotations among the genes identified by GWAS, (Table 2.6), indicates that the types of genes returned by this approach are predominantly regulators of enzymatic function rather than the enzymes themselves.

We hypothesise that this may be due to the insensitivity of the GWAS study used. Any difference in regulator function is likely to be amplified through the multiple enzymes they regulate, and therefore produce a larger phenotypic response. Additionally, whilst FBA simulates the complete knockout of reactions associated with each gene, the GWAS mapping population is unlikely to include individuals with completely non-functional enzymes; instead it features individuals with differently functional enzymes. Non-knockout mutations in enzymes are likely to lead to relatively subtle phenotype changes, and it is not clear that the phenotyping carried out in the GWAS study is sufficiently sensitive to identify these loci. The enzymes which are identified by the GWAS study tend to immediately proceed the production of the metabolite of interest, suggesting that indeed, relatively subtle, metabolically distant enzymes cannot be recovered. This is not to necessarily criticise the GWAS approach for failing to identify these subtle effects; at some point an effect must be acknowledged to be so subtle as to be effectively non-existent, and therefore perhaps uninteresting, at least from an engineering viewpoint.

We have discussed differences in the relative strengths of FBA and GWAS for identifying different types of genes. This explains why genes predicted by FBA might not also be identified by GWAS, however it does not explain why those genes identified by GWAS which are in the Arabidopsis model are not recovered. This can partially be attributed to the large number of false positives expected to be produced by linkage disequilibria in GWAS, but it must also be acknowledged to potentially be due to errors in the model structure, and the incomplete mapping of genes to reactions. Unfortunately it is not easy to distinguish between these causes.

Although FBA is better at predicting already expected genes involved in glucosinolate production than GWAS, it is not clear that it currently adds anything to the process of the identification of novel factors which is not already available through simple biological intuition. We initially hoped that FBA and GWAS could be used in conjunction to prioritise selection of genes for further study, however, we have seen that the different sensitivities of the approaches to different types of genes makes this impractical, at least in the particular example we have studied. Instead, in the future, as the quality of reaction network models continue to improve, it may perhaps be more appropriate to consider the use of

Table 2.6: The most common gene ontologies among the genes identified by GWAS as causing a glucosinolate phenotype indicate that this approach predominantly identifies regulators rather than enzymes. The most common associated location is the nucleus. Kinase, and protein and DNA binding are all more common recovered functions than catalytic activities. Regulation of transcription, and protein phosphorylation are among the most common annotated processes. This helps to explain the lack of agreement between GWAS and FBA approaches as the FBA model does not incorporate regulators.

(a) Location

| | Gene ontology | counts |
|---|---|---|
| 1 | nucleus | 853 |
| 2 | chloroplast | 384 |
| 3 | cytoplasm | 328 |
| 4 | plasma membrane | 316 |
| 5 | mitochondrion | 277 |
| 6 | chloroplast stroma | 268 |
| 7 | chloroplast envelope | 260 |
| 8 | membrane | 205 |
| 9 | chloroplast thylakoid membrane | 190 |
| 10 | vacuolar membrane | 126 |

(b) Function

| | Gene ontology | counts |
|---|---|---|
| 1 | ATP binding | 402 |
| 2 | molecular_function | 369 |
| 3 | protein kinase activity | 214 |
| 4 | protein serine/threonine kinase activity | 196 |
| 5 | protein binding | 181 |
| 6 | sequence-specific DNA binding transcription factor activity | 167 |
| 7 | catalytic activity | 137 |
| 8 | zinc ion binding | 118 |
| 9 | nucleotide binding | 115 |
| 10 | transferase activity, transferring phosphorus-containing groups | 100 |

(c) Processes

| | Gene ontology | counts |
|---|---|---|
| 1 | biological_process | 409 |
| 2 | regulation of transcription, DNA-dependent | 112 |
| 3 | oxidation-reduction process | 79 |
| 4 | protein phosphorylation | 75 |
| 5 | metabolic process | 71 |
| 6 | embryo development ending in seed dormancy | 49 |
| 7 | response to salt stress | 44 |
| 8 | positive regulation of transcription, DNA-dependent | 37 |
| 9 | signal transduction | 37 |
| 10 | transmembrane transport | 31 |

both approaches in parallel primarily in order to have increased sensitivity over all relevant types of genes, rather than for mutual validation.

## 2.4 Sulfur starvation comparison

### 2.4.1 Introduction

Sulfur is required for the growth of all organisms. In the environment, sulfur is most commonly found oxidised in the form of sulfate ($SO_4^{2-}$). Plants and micro-organisms are able to take up sulfate, and reduce it, prior to incorporation in organic molecules. Sulfate uptake and assimilation are tightly regulated according to plant demands for reduced sulfur (reviewed [216]).

As well as specialised, secondary metabolites, sulfur is required for the biosynthesis of essential metabolites such as cysteine and methionine. These exemplify the interconnectedness of sulfur metabolism with nitrogen and carbon. This connectedness means that sulfur starvation results in widespread metabolic responses in numerous metabolic pathways throughout metabolism, for example changes in nitrogen assimilation, photosynthesis, lipid breakdown, and auxin and jasmonate production [151, 152, 135, 87]. Widespread affects meant that the impact of sulfur stress is a potentially interesting target for FBA analysis using a genome scale model, rather than a more focused kinetic model, due to the large number of processes by which it could be affected.

Previous experimental studies have indicated that sulfur starvation results in a broadly biphasic response; the response to relatively mild sulfur starvation stress is very different to more severe stress (reviewed [80]).

Mild stress responses are generally specific to particular nutrients, for example the induction of specific high affinity transport systems. In the case of sulfur stress, the activity of high affinity sulfate uptake transporters (SULTR1;1 & SULTR1;2) is induced [248], and synthesis of glucosinolates is reduced [88]. During this initial response phase, resupply of nutrients can restore normal cellular functions and rescue the plant [80].

In contrast, severe nutrient stress leads to a more general emergency nutrient deficiency response which is shared between several nutrients [80], and which is characterised by an irreversible switch in the developmental program towards senescence and maturation [240]. Resupply with nutrients does not reverse this switch [80]. Nutrient depletion induced senescence is accompanied by reduced protein synthesis rates [225], and decreased photosynthetic assimilation of carbon. Although severe nutrient stress exhibits many common responses between nutrients, especially for the light reactions, Calvin-Benson cycle and photorespiration, they are not identical. For example long term sulfur starvation, causes the degradation of indole glucosinolates by nitrilase, in order to provide precursors for auxin biosynthesis [117].

We considered that this change in behaviour could be the consequence of switching between two distinct metabolic strategies, rather than a single increasingly severe response. To investigate this behaviour, we therefore incorporated the induction of sulfate transporters, and senescence into the model, as described in Methods. This allowed the FBA model to induce additional sulfate uptake transporter activity by paying a metabolic penalty cost related to the cost of the synthesis of more transporter proteins. Senescence was modelled by the creation of amino acids, with the constraint of reduced total flux through all reactions in the system. This is intended to reflect the reduction in catalytic enzyme activity, as in the FBA literature total flux through the system is often equated to the total amount of catalytic enzymes required [32].

By incorporating these two strategies into the optimisation framework, we could assess how the optimal behaviour for biomass production changed as sulfur stress increased, and see how closely the experimentally observed behaviour corresponded to the predicted optimal solution for the maximisation of biomass.

### 2.4.2 Nutrient response growth curve

We considered the behaviour of the induction, and senescence responses for the optimal production of biomass as environmental sulfur concentration was reduced (Figure 2.5, Figure 2.6). We manually explored the parameter space of the model in order to find a parameter set in which both induction and senescence were utilised in the induction and senescence models respectively, and in which we see that both induction and senescence strategies were used at some environmental sulfur concentration in the induction and senescence model.

Figure 2.5 shows that as environmental sulfur decreases, the amount of biomass producible decreases from an optimal level, of just over 0.8 arbitrary units, (at which point glucose availability is limiting for growth), down to zero biomass production at zero environmental sulfur for the induction, and base model, (in which neither senescence nor induction are permitted), and a small, non-zero value for the senescence, and induction and senescence models.

The growth curves for all four models approximate the classical saturation curve shape expected from experiments in which growth is increasingly adversely affected by increased nutrient stress. However, it is important to note that the x-axis in Figure 2.5 corresponds to environmental, rather than internal sulfur concentration curve, and that the saturation shape is purely a consequence of the Michaelis-Menten kinetics applied to the uptake reaction. Conversely the experimentally derived curve shape should also be seen when internal sulfur concentration is plotted on the x-axis. FBA is not well able to deal with metabolite concentrations, but we note that biomass production linearly corresponds to the sulfur uptake flux in the basal model, which we interpret such that biomass

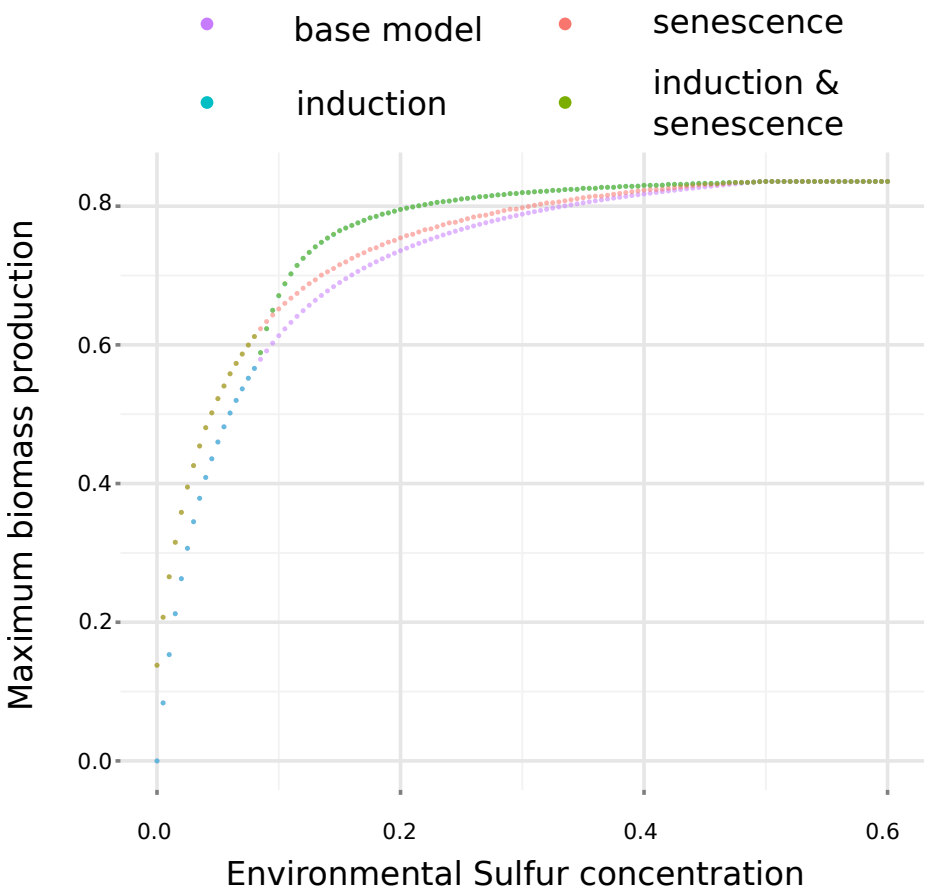



Figure 2.5: The effect of sulfur limitation on producible biomass in models of Arabidopsis metabolism. All units are arbitrary. In the base model, uptake of sulfur is related to the environmental concentration by Michaelis-Menten kinetics. In the induction model, the model is able to increase uptake of environmental sulfur through the production of more transporter. In the senescence model, import of amino acids is permitted, but permitible total flux through all reactions in the model is concurrently reduced. In the induction and senescence model, both strategies are available. This shows that the growth curves of all models appear similar to the classical effect of nutrient deprivation on growth, that under mild sulfur deprivation, induction of transporters is preferred over senescence, and that optimal behaviour appears to be a switch between induction and senescence, rather than their use in parallel at a given environmental sulfur concentration.

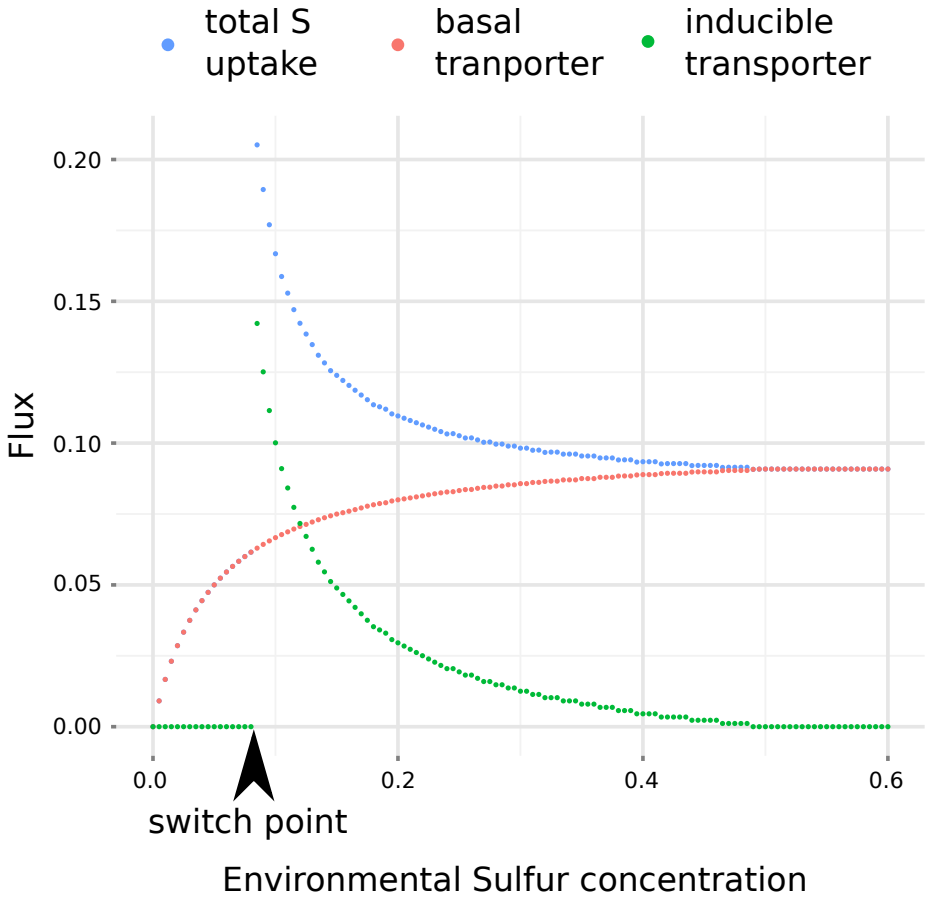



Figure 2.6: The effect of sulfur stress on flux through sulfur uptake reactions in the induction & senescence model. Basal transporter is flux through the un-inducible sulfate uptake reaction present in all models, total S uptake is the summed flux through the basal and inducible sulfate transported. The non-linear relationship between sulfur stress, and additional requirement for nutrients and total flux is shown by the nature of the increase in total S uptake flux at lower environmental sulfur concentrations above the level required at sufficient sulfur levels; the increase in total sulfur uptake above the level required at sufficient sulfur concentration is directly proportional to flux through the penalty reaction. The switch between induction and senescence behaviour, (highlighted by arrow), can be seen by the dramatic reduction in total S uptake at environmental sulfur concentration approximately equal to 0.08.

production could be predicted to be linearly related to internal sulfur concentration.

Although sulfate is taken up from the environment, and stored in the vacuole, it is exclusively reduced in the plastid, which it is transported into by SULTR3;1 [25]. In the presented model, concentration dependent kinetics were only incorporated into the (environmental) uptake transport step. Hence it is tempting to speculate that the saturation curve seen experimentally, with reduced internal nutrient concentration, could be a consequence of analogous kinetics in other transport or reaction steps reducing flux capacity as the concentration of internal metabolic pools decrease.

Figure 2.5 shows that as suggested experimentally, transporter induction is the optimal strategy under relatively small nutrient stress, in that by adopting this approach, the greatest amount of biomass flux can be achieved. This is only shown for one parameter set, however we were not able to find any model parameter set in which senescence was induced at relatively minor stress and induction then took place at more severe stresses. This is a consequence of the Michaelis-Menten kinetics used to calculate the penalty flux for activating the inducible sulfur flux, which means that as environmental sulfur concentration was decreased, the amount of transporter required to maintain a given uptake flux increased non-linearly. This non-linear increase in penalty is indicated in Figure 2.6, where the total amount of sulfur required is plotted, and increases exponentially as environmental sulfur decreases, as a consequence of flux through the penalty reaction. Consequently, if, (for a given parameter set), induction is not favourable at small nutrient stress, it cannot become favourable at higher stresses, hence we cannot see an optimal strategy of senescence at moderate stress, and induction at severe stress.

The rate of this non-linear increase in cost is a consequence of the arbitrary parameter $c$ in Equation 2.4, and the switching point between induction and senescence is dictated by the values of $c$ and $m$ (see Equation 2.5 in Methods). The parameter values used lead to somewhat unphysiological behaviour, (as indicated, for example, immediately adjacent to the labelled 'switching point', when environmental sulfur concentration is approximately 0.08 and almost all of the sulfur taken up is used to produce more uptake transporter), however the non-linear relationship will be be same for all positive values of $c$, and so this qualitative discussion holds.

Figure 2.5 and Figure 2.6 show that, as a consequence of this non-linear behaviour, at some environmental sulfur concentration (here approximately 0.08), the increase in flux through the penalty reaction becomes greater than the benefit of the increased availability of sulfur, and the optimal solution for the inducible model becomes the same as for the basal model, or the senescence model for the induction and senescence model, and hence induction ceases.

Interestingly, we were not able to find any parameter set for which the induction and senescence model was able to perform better than both the induction and

senescence models for some contiguous range of environmental sulfur concentrations. Consequently the induction and senescence model tracks whichever of the individual strategies performs better at a given sulfur concentration, switching between them, and at no evaluated point employing both behaviours concurrently (Figure 2.5).

This suggests that the optimal strategy for biomass production as sulfur availability decreases is indeed a switching behaviour between induction and senescence, as suggested by implication in the literature, rather than the additional recruitment of senescence to supplement induction of transporters. This is apparently an inevitable consequence of non-linear transporter kinetics with respect to substrate concentration, rather than the particular kinetic parameters seen experimentally.

This mutual exclusivity is likely to be because the induction of transporters requires a large number of metabolic products. Induction of uptake transporters therefore necessitates an increase not only in nutrient requirements (as shown in Figure 2.6, but also in total flux relative to the basal model, and so cannot be integrated effectively with the reduction in permitted total flux caused by senescence, leading to their exclusive use. We have not explicitly demonstrated that this switch is irreversible, such that additional supply of sulfur does not reverse the senescence phenotype, but given this result, it is intuitive that a system in which reduced total flux is available due to the use of senescence cannot recruit transporters to take advantage of additional nutrients supplied, without further (transiently) compromising biomass production. Hence, a greedy approach to optimisation, (in which every intermediate step must be better than the previous), cannot be expected to find the (globally) optimal solution of a return to transporter induction without senescence.

In summary, based upon this modelling, we conclude that optimal behaviour in response to nutrient stress is initial induction of high affinity transporters, followed by catabolism at increased stress, and switching between the exclusive use of induction and catabolism rather than their co-occurence. This result appears to be independent of the particular parameters used in the model, and is instead a consequence of the structure of the reaction network, and the transporter kinetics. Although these results are quite nice, we wanted to relate them more closely to experimental data, to assess more closely how well this predicted optimal behaviour relates to the experimentally determined behaviour of plants under different levels of sulfur stress.

Interestingly, we noticed that in many published studies, although the conditions imposed upon the plants are reported, it is not necessarily clear to what degree the organism is actually stressed. Different modelled flux distributions occur under the induced transporter, and senescent lifestyles, and the relative flux through these 'diagnostic reactions', could potentially be used to quantitatively determine the level of stress imposed on the plant in a particular study, similar to the approach taken by Cheung et al. [33] to estimate relative maintenance costs.

However for either of these ambitions to be realised, we wanted to confirm that the flux distributions seen under sulfur stress relate to experimental distributions. We therefore compared the predicted flux changes under sulfur stress to transcriptomic studies of sulfur starvation. As mentioned, the degree of stress imposed in these studies is often unclear, and we therefore fit the parameters in equations Equation 2.4 and Equation 2.5 to gene expression changes in the experimental data, so as not to impose an assumed induction or senescence based lifestyle upon the experimental system.

### 2.4.3 Comparing predictions to gene expression data

In order to validate the reaction flux predictions of the model, we compared it to published gene expression data under sulfur sufficient, and sulfur starved conditions.

As ever, the use of transcriptomic data is not ideal; beyond transcriptional regulation, many components of sulfur metabolism are additionally controlled by complex post-transcriptional regulation (reviewed in [216]). However, the available experimental data for sulfur starvation experiments was predominantly from micro-array studies. We therefore considered gene response to sulfur starvation as a binary up/down response, as this is likely more robust both to the limitations of the data, and the model.

FBA returns a single flux solution, however it is not necessarily the only optimal solution, and may not be representative. We therefore used FVA [132] in order to find the upper and lower flux limits for each reaction imposed by the objective function, and the reaction network structure. We derived the predicted response from the correlation of these upper and lower bounds as indicated in Table 2.12 in Methods.

Table 2.7 shows how closely the predictions correspond to previously published sulfur starvation experiments. Across three smaller datasets, the experimental data shows quite good agreement to the model predictions. However in the full dataset of Marayuma-Nakashita [136], there is only just over 50% agreement. As 'correct'/'incorrect' is only assessed at the level of increase, or decrease under low sulfur conditions, 50% correct is the level that would naively be expected by chance. Therefore the FBA model cannot be considered to be performing well on this dataset.

Interestingly, there is a difference in prediction performance between the genes which are discussed by Maruyama-Nakashita et al., and therefore included in the supplementary information of their paper, and their full dataset (Maruyama-Nakashita et al., 2006 *Supp. info* versus *All data* in Table 2.7). We believe that this is caused by two factors. Firstly, we speculate that Maruyama-Nakashita et al. preferentially picked genes for discussion for which they had some explanation, and therefore fit the existing biological knowledge which is incorporated in

Table 2.7: Comparison between predicted and experimental response to sulfur stress. Predictions were compared to three experiments. As the level of stress experienced by the plant is not known, for each experiment, the parameters $c$, and $M$ we fitted, so as to maximise the correct percentage. This effectively fits whether induction, catabolism, or neither strategy are used. The experimental data is from the papers [48, 151, 136]. Agreement to small datasets of curated, discussed data is comparatively good, compared to agreement in the full micorarray experiment. This is likely because the data selected for discussion in these papers exhibits a relatively large fold change, and makes intuitive biological sense, and is thus more likely to be captured by the reaction structure of the model.

| Data | # Correct | # Incorrect | % correct |
|---|---|---|---|
| D'Hooghe *et al.*, 2013 | 9 | 4 | 69.2 |
| Nikiforova *et al.*, 2003 | 25 | 7 | 78.1 |
| Maruyama-Nakashita *et al.*, 2006 *Supp. info* | 31 | 17 | 64.6 |
| Maruyama-Nakashita *et al.*, 2006 *All data* | 1279 | 1263 | 50.3 |

the Arabidopsis model, and which is therefore better captured by the predictions of the model than average.

Secondly, transcriptomic studies apply a ratio threshold above which a change in expression is considered to be real and discussed, and below which is ignored. The potential impact of this step on prediction quality is illustrated in Figure 2.7. It is not clear that the majority of transcripts 'truly', or importantly change under sulfur stress. The expression ratio change is small for the majority of genes, and could be due to experimental artefacts. Additionally, the modal expression ratio is not 1.0 as might be expected, but 1.2. It is not clear whether this shift is a true biological phenomenon, or an artefact of the experiment performed.

This modal shift in the experimental distribution means that almost all transcripts are considered to increase, whereas the predictions of the model are more equally divided between increased and decreased flux. In fact, given difference the distributions shown in Figure 2.7, numerical simulation suggests that the model prediction accuracy of 50.3% is better than the majority (89%) of randomly allocated predictions with the same distribution.

In the scores reported in Table 2.7, we considered all expression ratios as indicative of a change in expression. Figure 2.8 shows that the prediction quality increases as we impose a more extreme cutoff threshold for experimental changes in expression ratio, and that for very extreme fold changes, predictions are quite accurate. This is not to suggest that the lack of agreement is entirely caused by experimental error. The cutoff threshold continues to affect prediction quality far above the cutoff thresholds used in the literature which are usually in the order of 1.5-2.0x [48, 151, 136], (at which prediction accuracy is only 54.4%), suggesting that the model is truly better at predicting transcripts which vary

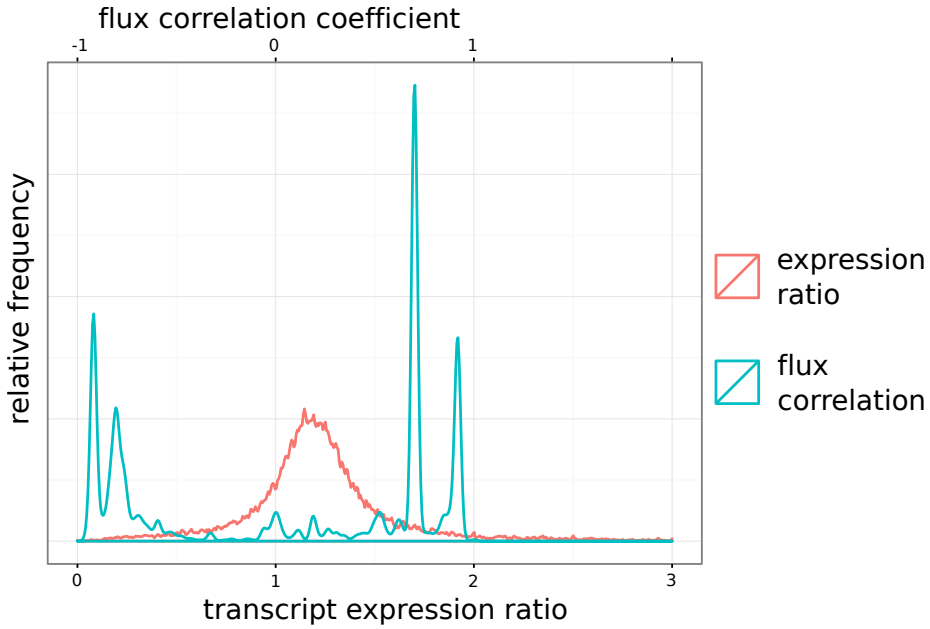



Figure 2.7: The distributions of predicted correlation coefficients, and expression ratios are very different. The majority of genes do not appear to vary much in response to sulfur stress, conversely the model incorrectly predicts that most reactions should either increase or decrease response to sulfur stress. The distribution of expression ratios is centred at approximately 1.2. It is not clear whether an expression ratio of 1.0, or 1.2 should be considered as the 'no change' value. By imposing a more extreme cutoff ratio, we are able to bystep this issue, as well as reducing experimental error, and only considering genes which are more likely to 'truly' vary in expression.

strongly (rather than weakly) in response sulfur starvation.

Although conditions can be found for which model performance is significantly improved, it is not necessarily clear from the predictions themselves which are more likely to be correct. We therefore attempted to find indicators of a predictions likely accuracy based only on the model output.

Figure 2.9 shows that there is no useful difference in the quality of the predictions based on the magnitude of the predicted correlation coefficient, although there is a slight tendency for predictions of a large, negative correlation between reaction flux and environmental sulfur to be more likely to be correct than others.

It is not expected that the quality of the reaction network model is equivalent across all regions of metabolism, and we thought that reaction location could potentially be used to weight prediction confidence. To integrate spatial information into our predictions, we divided the reaction network into clusters of contiguous reactions (the product of one is the substrate for the next) based on predictions of increased, or decreased flux under sulfur stress (Figure 2.10a).

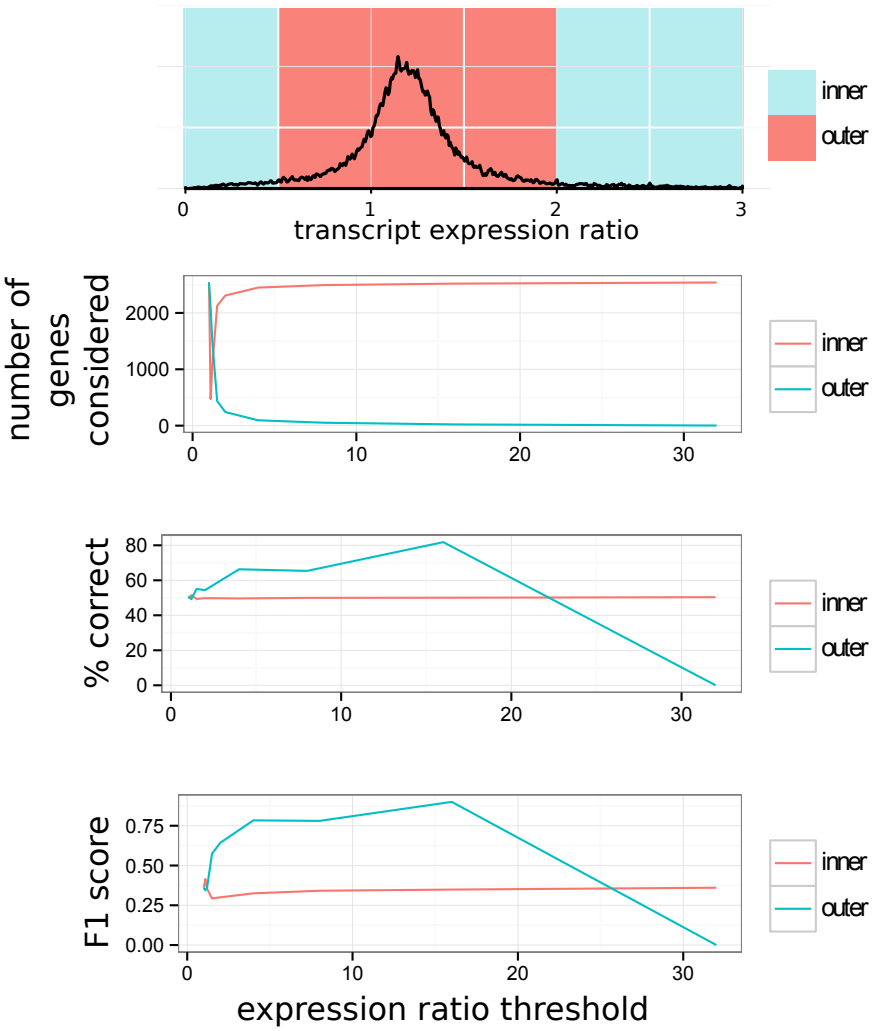



Figure 2.8: Prediction accuracy is greater for transcripts which vary more greatly under high and low sulfur conditions. When we varied the expression ratio threshold, and only considered genes whose expression changed by at least this value, we found that prediction accuracy, and that F1 score increased, although with a decrease in the number of genes which could be considered. The top panel illustrates 'inner' as those genes whose expression ratio is between the expression ratio threshold, and 1 over the expression ratio threshold. '% correct is the fraction of predictions which are correct, equivalent to 'precision' in Equation 2.6 calculation of the F1 score is described in Methods. Both % correct, and accuracy increase over the considered points, until no more genes are considered beyond a 32 fold cutoff threshold .

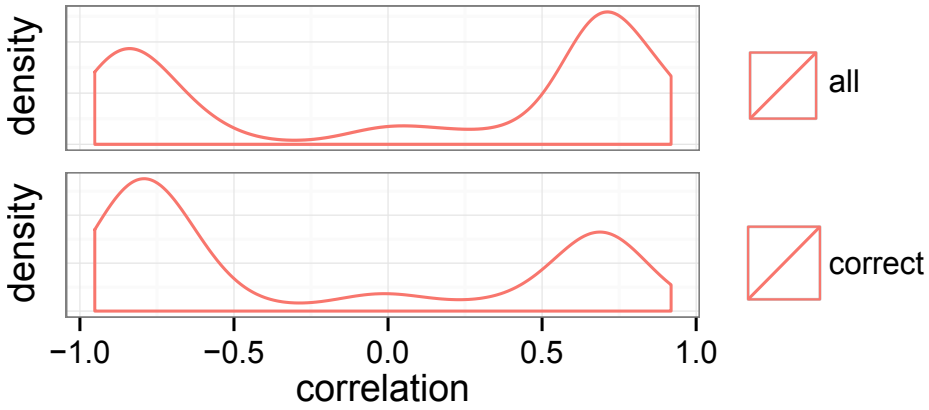


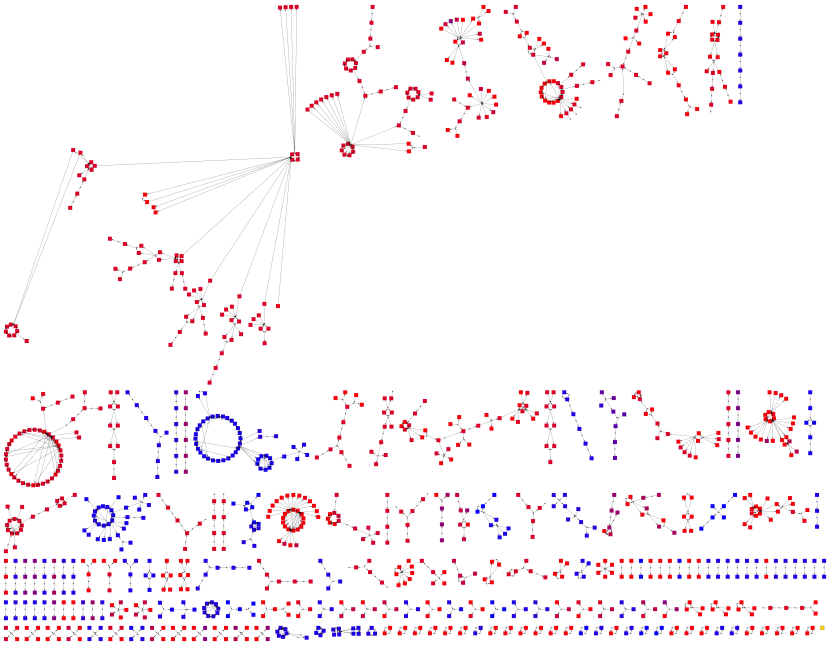

Figure 2.9: Correlation strength is not a good indicator of the reliability of a prediction. We see that there is little difference in the correlation coefficient distribution between reaction flux and environmental sulfur concentration for correct predictions, and all predictions, and therefore correlation strength cannot be used to distinguish a subset of higher quality predictions.

We then superimposed whether these predictions were in agreement or disagreement with experimental changes in the Maruyuma-Nakashita full dataset [136] onto these clusters (Figure 2.10b).

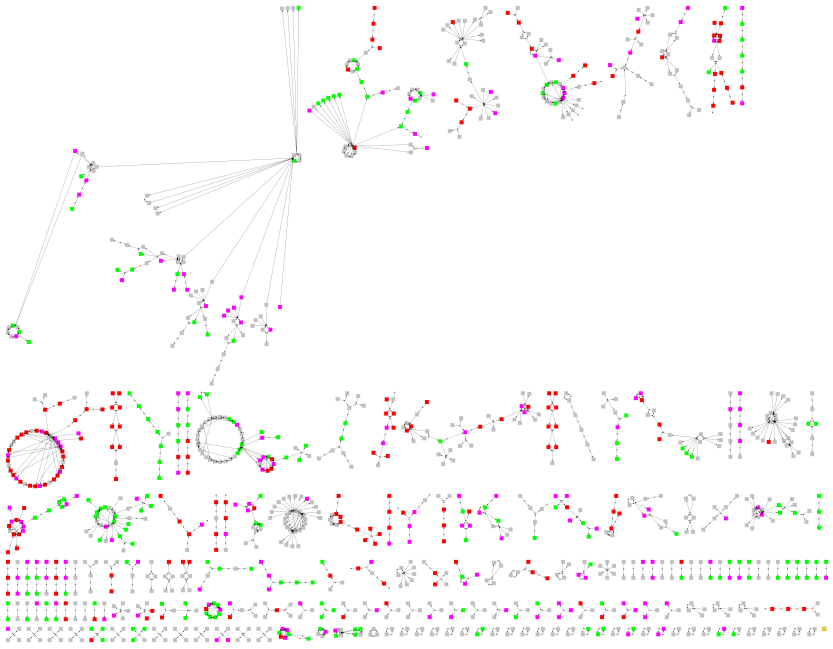
These clusters do tend to segregate correct and incorrect predictions, such that predictions within each cluster tend to be mostly correct, or mostly incorrect to an greater extent than would be expected if the correct and incorrect predictions were distributed randomly.

This distribution further validates the reaction network structure, as the structure causes flux through sets of reactions to covary in a way which is consistent with the observed experimental data. The dominant source of error is in the direction of the change under sulfur stress. The mistakes the model makes in comparison to the transcript data is therefore likely to be caused by either: 1.) the objective function, 2.) the implementation of induction, and senescence, or 3.) the assumption that induction and senescence are the plant response to sulfur starvation. However, although isolating the problem to these steps is good, in that it suggests that the time intensive process of reaction network building has not been a waste of time, it is not clear how these errors can be easily fixed under the FBA framework.

Figure 2.9 does suggest a potential method for the extension of the experimental data using this network structure information, without the need for the potentially false assumptions imposed. Within each cluster, a prediction of 'up' / 'down' can be assigned to the reactions which are not covered by the experiment (grey nodes), based on the consensus direction of the experimental data within the cluster. By integrating these changes as constraints on the flux bounds of



(a) Model predictions



(b) Prediction correctness

Figure 2.10: Network structure does relate to transcript behaviour. We broke up the reaction network into contiguous reaction clusters (such that the product of one reaction is the substrate for connected nodes) on the basis of predicted response to environmental sulfur (Figure 2.10a). Blue reaction nodes indicate a negative Pearson's correlation coefficient, red indicates positive. In Figure 2.10b, we superimposed a whether the production was correct or incorrect (green; correct, red; incorrect; purple; multiple linked genes which respond in different directions, grey; no mapped gene). We see that generally a linked reaction cluster is either correct or incorrect, indicating that transcripitional response does relate to network structure, and flux predictions, but the incorrect step is in the response of these clusters, although it is correct that the clusters should behave in the same way.

the Arabidopsis model as described in various papers, (reviewed [130]), this integration could also potentially provide insight as to the metabolic consequences of observed gene expression changes under sulfur starvation.

Overall in this section, we have seen that agreement between model predictions, and published gene expression data is reasonable, but is not very good. This is a consequence of the very different distributions of predicted, and experimental changes, however the extent to which these differences are artefacts of the experimental methodology is unclear.

Model derived errors are likely a consequence of the FBA optimisation assumption, and the implementation of induction, and senescence within the model, rather than the reaction network structure itself. The modelling approach we have taken is very crude. It is likely that the biomass equation should change under the 'senescent' lifestyle, and although FBA requires the assumption that metabolism is regulated so as to achieve some metabolic 'objective' (see chapter 1), it seems unlikely that this is well approximated by biomass production under senescence. Furthermore nutrient starvation leads to a reduction in photosynthesis, which causes photo-oxidation. This additional stress is not easily incorporated into the model, but is likely to lead to widespread metabolic changes in the experimental data. Under senescence, we imposed a general reduction in total flux, essentially penalising all reactions equally, however different reactions require different amounts of protein, depending on catalytic activity, and it is not clear that the data generation required for incorporating this idea into the model is practically accomplishable.

It is not necessarily easy to address these issues, and consequently the use of the model to further investigate the switching behaviour described in subsection 2.4.2 in response to sulfur starvation was reprioritised in favour of other uses for the genome scale reaction network.

## 2.5 Conclusion

In this chapter, we have further developed a previously published model of Arabidopsis metabolism. We have assessed the suitability of this model and FBA for the analysis of secondary metabolite phenotypes, and also used it for the investigation of plant response to nutrient deficiency stresses.

Genome scale models are important not only for use in predictive modelling, but also as data structures to integrate metabolic and genetic information. It is therefore important to continue to refine and expand these models as more information becomes available to keep them up to date, irrespective of immediate improvements to model FBA predictions. As such, although the further curation of the model has only led to slight improvements in predicted flux distribution relative to tracer experiment data, the improvement the prediction quality of

growth/no growth phenotypes indicates that the changes made are useful, and represents steps towards a 'correct' model of Arabidopsis metabolism.

We have consistently seen that FBA performs well in recapturing and providing additional insight to explain 'expected' behaviour. However, particularly in relation to the production of glucosinolates, it is difficult to claim that we have gained a useful, novel, insight in agreement with experimental data. In applying FBA to secondary metabolism, we have moved away from the traditional areas of strength for metabolic flux analysis, in order to assess its use for potentially more commercially interesting areas of metabolism. The quality of predictions is a function of the extent to which the behaviour of the studied phenomena is an emergent property of existing knowledge, integrated into the model, as opposed to unknown biological agents. Therefore although our findings suggest that the current state of plant genome scale models is not sufficiently high to be usefully applied to secondary metabolism, this is not expected to be the case forever. Interestingly, we have seen that FBA and GWAS approaches perform differently in relation to the types of genes which are recovered. This suggests that as plant models improve, FBA based approaches are likely to have a useful place in the suit of approaches for identifying the causes of complex, commercially interesting phenotypes.

Overall, we have seen abundant evidence that the underlying model used does correctly capture the reaction system. This was the case in the input output requirement predictions, high predicted flux correlation to MFA data in central metabolism, good recovery of expected glucosinolate mutants, and even reasonable predictions of the response to stress. Furthermore we have seen some evidence that errors are imposed on the model structure partially by the assumption of some optimality criteria required by flux balance analysis. Therefore in chapter 3, we use the same reaction network for elementary flux mode analysis, which does not impose these assumptions, and instead focuses more on the capabilities of the metabolic network.

## 2.6 Methods

### 2.6.1 Flux balance analysis

In this study, flux balance analysis was carried out using the COBRA toolbox (`https://opencobra.github.io/`) for MATLAB, and the Gurobi linear programming solver (`http://www.gurobi.com/`).

A metabolic network consisting of $C$ metabolites, and $R$ reactions can be represented by the stoichiometric matrix $S$, where element $S_{\{c,r\}}$ is the stoichiometric coefficient of metabolite $c$ in reaction $r$. Substrates have negative coefficients, products positive. $v_r$ represents flux through reaction $r$ For each reaction in $(r = 1, ..., R)$. Steady state flux solutions can be found by solving the equation

$$\sum_{r=1}^{R} S_{\{c,r\}} v_r = 0, \quad \forall c \in I \tag{2.1}$$

$$lb \leq v \leq ub$$

where $I$ is the set of all internal metabolites, and $lb$ and $ub$ are vectors of lower and upper bounds on permissible flux values.

Most plants take up nitrogen in the form of ammonia and nitrate. In this chapter, nitrogen uptake was constrained to be 50% nitrate, and 50% ammonium. Photorespiratory metabolism was simulated by additionally constraining the ratio of carboxygenic to oxygenic rubisco catalysed reactions to three. NADPH oxidation reactions in the cytosol, plastid, and mitochondria were constrained to the ratio 1:1:1.

A 'maintenance cost' representing miscellaneous energy requirements of the cell is commonly applied, by the additional constraint

$$v_{ATPase} = m \tag{2.2}$$

where $m$ is constant, and $v_{ATPase}$ is flux through the ATPase reaction

$$\text{ATP} \longrightarrow \text{ADP} + \text{Pi}.$$

In flux balance analysis, degenerate solution space is reduced by additionally imposing an objective function. Here the objective function was to maximise: $v_b$ subject to Equation 2.1, and Equation 2.2, where $v_b$ is flux through some 'biomass reaction', an export reaction, in which individual biomass components are exported from the model, representing the production of biomass in the organism.

For each previously published model we assessed, we used the incorporated biomass model.

The exported components, and the required ratios used in the modified Cheung model are shown in Table 2.8, and were predominantly derived from the equation used in [33], and supplemented by the conclusions of manual curation as described in subsection 2.2.1 & subsection 2.2.2.

In order to find blocked reactions, rather then maximising flux through the biomass reaction, flux through each reaction was maximised and minimised. Reactions for which maximum and minimum flux were both zero were considered to be blocked.

Because FBA does not guarantee the existence of a single solution, flux variability analysis (FVA) was used as previously described [132]. As such, after the primary objective was initially optimised, the optimal value was imposed as an additional constraint, and flux through each individual reaction was maximised and minimised in order to find its permissible bounds.

Table 2.8: The metabolic components of the biomass equation.

| Metabolite | Ratio |
|---|---|
| Starch | 0.73737832 |
| Cellulose | 5.24035317 |
| Xylan | 1.13494168 |
| Fatty Acid | 0.77115239 |
| Glycerol | 0.35252681 |
| 4-aminobutanoate | 0.09635973 |
| Fumarate | 0.00569181 |
| Sucrose | 0.20702776 |
| Citrate | 0.26372076 |
| Malate | 0.42965891 |
| TYR | 0.16540407 |
| GLU | 0.29439008 |
| LYS | 0.25644188 |
| VAL | 0.31204192 |
| PHE | 0.23737933 |
| GLN | 0.29439008 |
| THR | 0.20656579 |
| MET | 0.10868391 |
| SER | 0.36452647 |
| GLY | 0.20656579 |
| HIS | 0.07877583 |
| LEU | 0.35940421 |
| ASP | 0.24423117 |
| ILE | 0.16493083 |
| ALA | 0.45610963 |
| ASN | 0.24423117 |
| ARG | 0.24870892 |
| ASP | 0.09364807 |
| Potassium | 1.80000000 |
| Calcium | 0.88000000 |
| Magnesium | 0.58000000 |

(a) Published Cheung 2013 model.

| Metabolite | Ratio |
|---|---|
| **growth no growth** | |
| ascorbate | $1e^{-6}$ |
| beta-alanine | $1e^{-6}$ |
| biotin | $1e^{-6}$ |
| chlorophyll-A | $1e^{-6}$ |
| cholesterol | $1e^{-6}$ |
| glutathione | $1e^{-6}$ |
| lipid-IV-A | $1e^{-6}$ |
| NAD | $1e^{-6}$ |
| pantothenate | $1e^{-6}$ |
| plastoquinol | $1e^{-6}$ |
| pPRO | $1e^{-6}$ |
| putrescine | $1e^{-6}$ |
| thiamine pyrophosphate | $1e^{-6}$ |
| tetrahydrofolate | $1e^{-6}$ |
| tocopherol | $1e^{-6}$ |
| **blocked reactions** | |
| cdp-ethanolamine | $1e^{-6}$ |
| pTRP | $1e^{-6}$ |
| coumarin | $1e^{-6}$ |
| dhurrin | $1e^{-6}$ |
| gibberellin $A_{110}$ | $1e^{-6}$ |
| gibberellin $A_{97}$ | $1e^{-6}$ |
| siroheme | $1e^{-6}$ |
| heme | $1e^{-6}$ |
| Beta-alanine betaine | $1e^{-6}$ |
| sinapaldehyde glucoside | $1e^{-6}$ |
| syringin | $1e^{-6}$ |
| quercetin 3 3' 4' 7-tetrasulfate | $1e^{-6}$ |
| benzyl-isothiocyanate | $1e^{-6}$ |
| 1 3 5-trimethoxybenzene | $1e^{-6}$ |
| thio-molybdenum cofactor | $1e^{-6}$ |
| ayapin | $1e^{-6}$ |
| coniferaldehyde glucoside | $1e^{-6}$ |
| demethylmenaquinone-13 | $1e^{-6}$ |
| all-trans-4 4'-diapophytofluene | $1e^{-6}$ |
| all-trans-dodecaprenyl diphosphate | $1e^{-6}$ |
| L-methionine-(S)-S-oxide | $1e^{-6}$ |
| cyclic AMP | $1e^{-6}$ |
| all-trans-undecaprenyl diphosphate | $1e^{-6}$ |
| dammarenediol II | $1e^{-6}$ |
| 3-methylsulfinylpropyl-glucosinolate | $1e^{-6}$ |
| 8-methylthiooctyl-glucosinolate | $1e^{-6}$ |
| 5-methylthiopentyl glucosinolate | $1e^{-6}$ |
| 7-methylthioheptyl-glucosinolate | $1e^{-6}$ |
| 6-methylthiohexylglucosinolate | $1e^{-6}$ |
| 4-methylsulfinylbutyl-glucosinolate | $1e^{-6}$ |
| 4-methylsulfinylbutyl-glucosinolate | $1e^{-6}$ |
| indolylmethyl-glucosinolate | $1e^{-6}$ |

(b) Added during model curation due to either the growth/no growth prediction analysis, or the blocked reaction analysis.

### 2.6.2 Comparison of metabolic flux analysis to flux balance analysis

Metabolic flux analysis (MFA) is a method for the determination of metabolic pathway fluxes based on the experimentally determined distribution of a labelled element among metabolites. Although it is based much more closely on experimental data than FBA, it relies on the imposition of a conceptual metabolic model of the intracellular reactions, and fluxes are predicted for this model. For comparison between metabolic flux analysis, and flux balance analysis, MFA reaction fluxes were mapped onto the FBA model via the grouping of FBA reactions, the fluxes of which were combined, and compared to the MFA prediction. The flux range of grouped FBA reactions was calculated by FVA of the summed flux though the grouped reactions. Mapping to metabolic flux data in [33, 139, 243, 244] was as described in [33, 243], otherwise the mapping is as shown in Table 2.9.

### 2.6.3 Mapping genes to reactions

We mapped genes to model reactions, by incorporating the information in the TAIR database ([91], `https://www.arabidopsis.org/`) and Biocyc databases ([253], `http://biocyc.org/`) into the curated Arabidopsis model. This resulted in the association of 4,037 unique genes to 1,812 reactions. Genes are often associated with to multiple reactions, and *vice versa*. This mapping was purely automated, and is not considered to be of particularly high quality, for example these databases generally do not include compartment specific information, and genes associated with reactions which occur in multiple compartments were mapped to all instances of the reaction.

### 2.6.4 Identifying genes & reactions predicted to affect glucosinolate production

To identify the genes which are predicted to affect glucosinolate production, we carried out FBA. After removing all glucosinolates from the biomass equation, we calculated maximum biomass production ($b_{wt}$), and maximum production of the target glucosinolate, ($t_{wt}$), in the *wild-type* plant for each glucosinolate in the model (Table 2.14).

For 'reaction knockout'; for each target glucosinolate ($t$), we sequentially applied the constraint $v_i = 0 \quad \forall (i = 1, ..., R)$ to simulate the knockout of every reaction singly. Alternatively, for 'gene knockout'; we simulated the knockout of each gene, by simultaneously constraining flux through all reactions associated with a given gene to zero. This is expected to result in a relatively large number of false positives, as 1.) not all genes associated with a reaction are expected to be essential for that reaction, 2.) due to the lack of compartment information

Table 2.9: MFA to FBA mapping for Chen 2013, and Szecowka 2013 datasets.

| MFA identity | Model reactions mapped to |
|---|---|
| **Chen 2013** | |
| Sucex $\longrightarrow$ SucCellular | $SUC_{tx}$ |
| Suc $\Longleftrightarrow$ UDP-Glc+Fru | -SUCROSE-SYNTHASE-$RXN_c$ |
| Suc $\longrightarrow$ Glc + Fru | RXN-$1461_c$ + RXN-$1461_v$ |
| F6P $\Longleftrightarrow$M6P | -MANNPISOM-$RXN_c$ |
| M6P $\Longleftrightarrow$ M1P | PHOSMANMUT-$RXN_c$ |
| M1P $\longrightarrow$ GDP-Man | RXN4FS-$12_c$ + 2.7.7.13-$RXN_c$ - MANNPGUANYLTRANGDP-$RXN_c$ |
| GDP-Man $\longrightarrow$ GDP-fuc | $\frac{\text{GDPMANDEHYDRA-RXN}_c+\text{1.1.1.271-RXN}_c}{2}$ |
| UDP-Gal $\Longleftrightarrow$ UDP-Glc | $-$UDPGLUCEPIM-$RXN_c$ $-$ GALACTURIDYLTRANS-$RXN_c$ |
| G1P $\longrightarrow$ UDP-Glc | GALACTURIDYLYLTRANS-$RXN_c$ + RXN-$11505_c$ + GLUC1PURIDYLTRANS-$RXN_c$ |
| G1P $\Longleftrightarrow$ G6P | PHOSPHOGLUCMUT-$RXN_c$ + PHOSPHOGLUCMUT-$RXN_p$ |
| UDP-Glc $\longrightarrow$ UDP-Rha | RXN-$5482_c$ |
| UDP-GlcA $\Longleftrightarrow$ UDP-GalA | UDPGLUCEPIM-$RXN_c$ + GALACTURIDYLYLTRANS-$RXN_c$ |
| G6P $\longrightarrow$ 6PG | $\frac{\text{GLU6PDEHYDROG-RXN}_c+\text{6PGLUCONOLACT-RXN}_c}{2} + \frac{\text{GLU6PDEHYDROG-RXN}_c+\text{6PGLUCONOLACT-RXN}_c}{2}$ |
| 6PG $\longrightarrow$ OPPP | 6PGLUCONDEHYDROG-$RXN_c$ + 6PGLUCONDEHYDROG-$RXN_p$ |
| Glc $\longrightarrow$ Glc_ex | -$GLC_{tx}$ |
| G1P $\longrightarrow$ Starch | $\frac{-\text{RXN-10770 + GLUC1PADENYLTRANS-RXN}_p+\text{GLYCOGENSYN-RXN}_p}{2}$ |
| Fru $\longrightarrow$ F6P | FRUCTOKINASE-$RXN_c$ |
| Glc $\longrightarrow$ G6P | GLUCOKIN-$RXN_c$ + GLUCOKIN-$RXN_p$ |
| UDP-Xyl $\longrightarrow$ Xyl | RXN-9104 |
| **Szecowka 2013** | |
| DHAP $\Longleftrightarrow$ FBP[p] | -F16ALDOLASE-$RXN_p$ |
| FBP[p] $\Longleftrightarrow$ F6P[p] | F16BDEPHOS-$RXN_p$ - 6PFRUCTPHOS-$RXN_p$ |
| F6P[p] $\Longleftrightarrow$ G6P[p] | -PGLUCISOM-$RXN_p$ |
| G6P[p] $\Longleftrightarrow$ G1P[p] | PHOSPHOGLUCMUT-$RXN_p$ |
| G1P[p] $\Longleftrightarrow$ ADPG | GLUC1PADENYLTRANS-$RXN_p$ - RXN-$10770_p$ |
| DHAP $\Longleftrightarrow$ FBP[c] | -F16ALDOLASE-$RXN_c$ |
| FBP[c] $\Longleftrightarrow$ F6P[c] | -6PFRUCTPHOS-$RXN_c$ - 2.7.1.90-$RXN_c$ + F16BDEPHOS-$RXN_c$ |
| F6P[c] $\Longleftrightarrow$ G6P[c] | -PGLUCISOM-$RXN_c$ |
| G6P[c] $\Longleftrightarrow$ G1P[c] | PHOSPHOGLUCMUT-$RXN_c$ |
| G1P[c]$\Longleftrightarrow$ UDPG | -GALACTURIDYLYLTRANS-$RXN_c$ + GLUC1PURIDYLTRANS-$RXN_c$ - RXN-$11505_c$ |

in reaction:gene databases, we often associate a gene to every instances of the catalysed reaction across all compartments.

In knockouts for which

$$t_{KO} < t_{wt} - \frac{t_{wt}}{1000}$$

meaning that the potential to produce the glucosinolate of interest is reduced, for reaction knockout, we report all genes associated with the reaction, or for gene knockout we report the gene identity.

These methods also return knockouts which would not be considered glucosinolate mutants, because the effect is much more general, for example by preventing any non-zero steady state flux through any reactions. We therefore also considered a variant in which reactions/genes were only reported if biomass, ($b$), production was not catastrophically effected in the knockout, that is if

$$b_{KO} > \frac{b_{wt}}{10}.$$

### 2.6.5 Modelling sulfur limitation

To model sulfur limitation simply, we used the Michaelis-Menten equation Equation 2.3 to relate the the maximum allowed flux through the sulfate uptake reaction to the environmental sulfur concentration.

$$v = \frac{V_{max} \cdot [S_{env}]}{K_m + [S_{env}]} \tag{2.3}$$

We arbitrarily set the Michaelis-Menten constant, $K_m = 0.05$, and $V_{max} = 1000$, $v$ is the calculated upper limit for permissible flux through this basal sulfate transporter. We modified the level of sulfur stress by varying the concentration of environmental sulfur, $[S_{env}]$.

To model the response to decreased sulfur availability, we assumed that sulfur starvation can lead to two physiological responses; induction of sulfate transporters to increase uptake from the environment, and senescence, leading to catabolism of complex molecules to recycle of sulfur, and a reduction in total catalytic enzyme capacity.

We added an 'inducible' sulfate uptake reaction,

$$3\mathrm{H}^+_{pumped} \xrightarrow{v_{ind}} \mathrm{SO}_4^{2-}$$

which allows the uptake of sulfate from the environment, powered by a proton gradient, and a cost, or penalty reaction in which tRNA molecules charged with amino acids, and pseudo metabolites representing energy costs, are consumed. This represents the increased protein required for the production of additional transporters (Table 2.10).

Table 2.10: The $v_{penalty}$ reaction equation consumes charged-tRNA molecules, and energy associated with protein production. The ratios of amino acids are based on relative amino acid frequencies in the FASTA sequences in the inducible Sulfate transporters SULTR1;1, and SULTR1;2.

| Substrates | | | Products | |
|---|---|---|---|---|
| 0.058 | Charged-ALA-tRNAs[c] | | 0.058 | ALA-tRNAs[c] |
| 0.023 | Charged-ARG-tRNAs[c] | | 0.023 | ARG-tRNAs[c] |
| 0.020 | Charged-ASN-tRNAs[c] | | 0.020 | ASN-tRNAs[c] |
| 0.032 | Charged-ASP-tRNAs[c] | | 0.032 | ASP-tRNAs[c] |
| 0.007 | Charged-CYS-tRNAs[c] | | 0.007 | CYS-tRNAs[c] |
| 0.022 | Charged-GLN-tRNAs[c] | | 0.022 | GLN-tRNAs[c] |
| 0.018 | Charged-GLT-tRNAs[c] | | 0.018 | GLT-tRNAs[c] |
| 0.045 | Charged-GLY-tRNAs[c] | | 0.045 | GLY-tRNAs[c] |
| 0.013 | Charged-HIS-tRNAs[c] | | 0.013 | HIS-tRNAs[c] |
| 0.063 | Charged-ILE-tRNAs[c] | | 0.063 | ILE-tRNAs[c] |
| 0.066 | Charged-LEU-tRNAs[c] | $\xrightarrow{v_{penalty}}$ | 0.066 | LEU-tRNAs[c] |
| 0.038 | Charged-LYS-tRNAs[c] | | 0.038 | LYS-tRNAs[c] |
| 0.018 | Charged-MET-tRNAs[c] | | 0.018 | MET-tRNAs[c] |
| 0.044 | Charged-PHE-tRNAs[c] | | 0.044 | PHE-tRNAs[c] |
| 0.032 | Charged-PRO-tRNAs[c] | | 0.032 | PRO-tRNAs[c] |
| 0.045 | Charged-SER-tRNAs[c] | | 0.045 | SER-tRNAs[c] |
| 0.035 | Charged-THR-tRNAs[c] | | 0.035 | THR-tRNAs[c] |
| 0.005 | Charged-TRP-tRNAs[c] | | 0.005 | TRP-tRNAs[c] |
| 0.018 | Charged-TYR-tRNAs[c] | | 0.018 | TYR-tRNAs[c] |
| 0.051 | Charged-VAL-tRNAs[c] | | 0.051 | VAL-tRNAs[c] |
| 0.653 | Protein-polymerisation-cost | | | |
| 0.653 | Protein-processing-cost | | | |
| 0.653 | Protein-tranlocation-cost | | | |

Flux through $v_{ind}$ and $v_{penalty}$ reactions are related based upon the Michaelis-Menten equation (Equation 2.3). We assumed that flux through $v_{penalty}$ is proportional to enzyme concentration, and therefore to $V_{max}$ in Equation 2.3, giving

$$v_{penalty} = \frac{v_{ind} \cdot (K_m + [S_{env}])}{[S_{env}] \cdot c} \tag{2.4}$$

in which $K_m$ is the Michaelis-Menten constant, $[S_{env}]$ is the concentration of environmental sulfur available for uptake by the plant (arbitrary units), and $c$ is some constant relating flux through $v_{penalty}$ to $V_{max}$, scaling the severity of the cost of inducing transporters. We set $K_m = 0.05$, the same as that for the basal, uninducible sulfate transporter.

To model potential catabolism of proteins to enable the recycling of sulfur, we added a reaction, $v_{cat}$, which allows the import of amino acids in the ratios shown in Table 2.11.

The penalty imposed for using the $v_{cat}$ reaction to produce amino acids is that allowable total flux through the system was reduced, proportional to flux through

Table 2.11: Amino acids and stoichiometries produced by the $v_{cat}$ reaction. Relative ratios of amino acids are proportional to their frequency in Arabidopsis, based on [119].

| $v_{cat}$ | products |
|---|---|
| 4.0 | ASP[c] |
| 0.5 | CYS[c] |
| 7.0 | GLT[c] |
| 1.0 | ASN[c] |
| 3.5 | SER[c] |
| 5.0 | GLN[c] |
| 0.5 | HIS[c] |
| 1.5 | GLY[c] |
| 2.5 | THR[c] |
| 0.05 | ARG[c] |
| 2.5 | THR[c] |
| 0.05 | ARG[c] |
| 2.0 | ALA[c] |
| 0.05 | TYR[c] |
| 0.05 | TRP[c] |
| 0.05 | MET[c] |
| 0.2 | VAL[c] |
| 0.05 | PRO[c] |
| 0.1 | PHE[c] |
| 0.05 | ILE[c] |
| 0.05 | LEU[c] |
| 0.5 | LYS[c] |

$v_{cat}$, reflecting a reduction in available enzymes. This was implemented by adding the constraint

$$Mv_{cat} + \sum_{i=1}^{N} v_i^{-S} \leq \sum_{i=1}^{N} v_i^{+S} \tag{2.5}$$

where $v_{cat}$ is flux through the $v_{cat}$ reaction, $v_i^{-S}$ is flux through reaction $i$ under test, low sulfur conditions, $v_i^{+S}$ is flux through reaction $i$ under control, sulfur replete conditions, and M is some arbitrary scaling constant.

The arbitrary constants $c$, and $M$, were manually varied to explore the behaviour or biomass production at different environmental sulfate concentrations, and set so as to allow both catabolism and induction to be utilised over the range of tested sulfate concentrations.

### 2.6.6 Comparison to transcript data

For comparison to transcript data, parameters $M$ and $c$ were optimised so as to minimise the number of incorrect predictions using the `scipy.basinhopping` algorithm, as it was not clear whether induction, catabolism, or both were occurring in the experimental data, and by doing this, the use of induction or catabolism can be fitted to the data.

Flux balance analysis often returns only one example of a number of optimal flux distributions, we therefore used flux variability analysis (FVA) [132], in which after the initial optimisation step, this is added as an additional constraint, and flux through each reaction is maximised and minimised to calculate the upper and lower flux bounds permitted by the objective function, and reaction network structure. Reaction response can be a relatively complex, non-linear function with respect to environmental sulfur concentration. To simplify this to a single number for each boundary, we calculated the correlation coefficient between the bound, and environmental sulfur concentration.

We compared these predicted bounds to the transcript expression ratio under sulfur stressed, and sufficient conditions from previously published studies ([48, 151, 136]).

The model predicts changes in the lower, and upper bound, therefore each reaction is associated with two predictions, in comparison the transcrptiomic data give only a single number per gene product. In order to map predictions of the changing flux boundaries to the experimental data, and score predictions as either correct, incorrect, or uninformative, we employed Table 2.12 as a lookup-table to score predictions.

We considered reactions for which the bounds change in an opposite manner as uninformative, but did compare predictions for which only one bound changes, as for many reactions redundancy in metabolic pathways means that the lower flux bound is unchanging, and zero across sulfur concentrations.

We considered predictions qualitatively rather than quantitatively due to the simplifying use of correlation coefficient to determine the models prediction, and the often poor correlation between transcript abundance and catalytic activity.

#### 2.6.6.1 F1 score

The F1 score is a measure of a predictive model's performance. It considers the 'precision' of the predictions (the number of correct positive predictions divided by the total number of positive predictions) and 'recall', the fraction of true values which are correctly recovered (number of correct positive predictions over true positive samples). 'Positive' predictions were considered to be either

Table 2.12: Table for scoring flux boundary predictions relative to experimental expression ratios. Predicted correlation is correlation between the predicted boundary, and environmental sulfur available, $\frac{-S}{+S}$ 'Expression Ratio' is the experimental transcript expression ration in low versus high sulfur conditions, ✓ indicates a correct prediction, ✗ indicates an incorrect prediction, – indicated an uninformative prediction. Cases where multiple gene identifiers map to a reaction or *vice versa*, and the genes/reaction did not behave in a consistent manner were ignored.

| Predicted Correlation | | $\frac{-S}{+S}$ Expression Ratio | |
|---|---|---|---|
| min | max | $< 1$ | $> 1$ |
| +ve | +ve | ✓ | ✗ |
| +ve | −ve | – | – |
| +ve | 0 | ✓ | ✗ |
| 0 | +ve | ✓ | ✗ |
| 0 | 0 | – | – |
| 0 | −ve | ✗ | ✓ |
| −ve | 0 | ✗ | ✓ |
| −ve | +ve | – | – |
| −ve | −ve | ✗ | ✓ |

predicted increase, or decrease in flux in response to sulfur stress, depending upon which response was rarer. The $F_1$ score is defined as

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.6}$$

$$precision = \frac{|u \cap v|}{|v|} \tag{2.7}$$

$$recall = \frac{|u \cap v|}{|u|}, \tag{2.8}$$

where $v$ is the set of rare predictions, and $u$ is the set of genes whose expression ratio indicates the same response.

# 2.7 Appendix

Table 2.13: Genes known or predicted to be involved in glucosinolate metabolism and regulation, reproduced from Chan et al. [28]. AGI shows the Arabidosis Genome Initiative identifier for each gene; Pathway shows the particular part of glucosinolate metabolism in which the gene is predicted to function; Evidence shows the experimental evidence (Genetic or Biochemical) or sequence evidence based on homology to a validated glucosinolate gene (Homology).

| AGI | Gene name/Description | Pathway | Evidence |
|---|---|---|---|
| AT5G25980 | TGG2 | GLS Breakdown | Biochem |
| AT5G26000 | TGG1 | GLS Breakdown | Biochem |
| AT5G48375 | TGG3 | GLS Breakdown | Biochem |
| AT1G47600 | TGG4 | GLS Breakdown | Biochem |
| AT1G51470 | TGG5 | GLS Breakdown | Biochem |
| AT1G51490 | TGG6 | GLS Breakdown | Biochem |
| AT3G09260 | PYK10 | GLS Breakdown | Biochem |
| AT1G66280 | glycosyl hydrolase family 1 protein | GLS Breakdown | Homology |
| AT2G44490 | PEN2 | GLS Breakdown | Biochem |
| AT1G52400 | BGL1 | GLS Breakdown | Homology |
| AT1G54040 | ESP | GLS Breakdown | Biochem |
| AT1G54045 | | GLS Breakdown | Biochem |
| AT3G16390 | AtNSP3 | GLS Breakdown | Biochem |
| AT3G16400 | AtNSP1 | GLS Breakdown | Biochem |
| AT2G33070 | AtNSP2 | GLS Breakdown | Biochem |
| AT3G16410 | AtNSP4 | GLS Breakdown | Biochem |
| AT5G48180 | AtNSP5 | GLS Breakdown | Biochem |
| AT3G07720 | | GLS Breakdown | Homology |
| AT3G14210 | ESM1 | GLS Breakdown | Biochem |
| AT1G54010 | AGG1/ESM2 | GLS Breakdown | Biochem |
| AT1G54020 | myrosinase-associated protein | GLS Breakdown | Homology |
| AT1G54000 | myrosinase-associated protein | GLS Breakdown | Homology |
| AT3G14220 | GDSL-motif lipase/ hydrolase family protein | GLS Breakdown | Homology |
| AT1G54030 | GDSL-motif lipase family protein | GLS Breakdown | Homology |
| AT1G52030 | MBP2 | GLS Breakdown | Biochem |
| AT1G52040 | MBP1 | GLS Breakdown | Biochem |
| AT2G39330 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT2G39310 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT3G16470 | JR1 | GLS Breakdown | Homology |
| AT3G16450 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT3G21380 | myrosinases binding protein like protein | GLS Breakdown | Homology |
| AT3G16440 | ATMLP-300B | GLS Breakdown | Homology |
| AT3G16460 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT2G25980 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT1G33790 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT1G52100 | jacalin lectin family protein | GLS Breakdown | Homology |

| | | | |
|---|---|---|---|
| AT1G60110 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT1G60095 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT1G57570 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT5G35940 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT5G35950 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT1G60130 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT1G52000 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT3G16420 | PBP1 | GLS Breakdown | Biochem |
| AT3G16430 | jacalin lectin family protein | GLS Breakdown | Homology |
| AT1G16410 | CYP79F1 | Aliphatic Glucosinolate | Biochem |
| AT1G16400 | CYP79F2 | Aliphatic Glucosinolate | Biochem |
| AT1G18590 | ATST5C | Aliphatic Glucosinolate | Biochem |
| AT1G24100 | UGT74B1 | Aliphatic Glucosinolate | Biochem |
| AT1G31180 | IPMDH | Aliphatic Glucosinolate | Homology |
| AT1G62540 | GSOX2 | Aliphatic Glucosinolate | Biochem |
| AT1G62560 | GSOX3 | Aliphatic Glucosinolate | Biochem |
| AT1G62570 | GSOX4 | Aliphatic Glucosinolate | Biochem |
| AT1G65860 | GSOX1 | Aliphatic Glucosinolate | Biochem |
| AT1G65880 | BZO1 | Aliphatic Glucosinolate | Biochem |
| AT1G12140 | GSOX5 | Aliphatic Glucosinolate | Biochem |
| AT1G74090 | ATST5B | Aliphatic Glucosinolate | Biochem |
| AT2G20610 | C-S LYASE | Aliphatic Glucosinolate | Homology |
| AT2G25450 | GS-OH | Aliphatic Glucosinolate | Genetic |
| AT2G31790 | UGT | Aliphatic Glucosinolate | Biochem |
| AT2G43100 | Aconitase | Aliphatic Glucosinolate | Homology |
| AT3G03190 | AtGSTF11 | Aliphatic Glucosinolate | Guess |
| AT3G19710 | BCAT4 | Aliphatic Glucosinolate | Biochem |
| AT3G49680 | BCAT3 | Aliphatic Glucosinolate | Guess |
| AT3G58990 | Aconitase | Aliphatic Glucosinolate | Homology |
| AT4G03050 | AOP3 | Aliphatic Glucosinolate | Biochem |
| AT4G03060 | AOP2 | Aliphatic Glucosinolate | Biochem |
| AT4G13770 | CYP83A1 | Aliphatic Glucosinolate | Biochem |
| AT4G12030 | Bile Acid Transporter | Aliphatic Glucosinolate | Guess |
| AT4G13430 | Aconitase | Aliphatic Glucosinolate | Homology |
| AT4G13770 | CYP83A1 | Aliphatic Glucosinolate | Biochem |
| AT5G23010 | MAM1 | Aliphatic Glucosinolate | Biochem |
| AT5G23020 | MAM3 | Aliphatic Glucosinolate | Biochem |
| AT5G07460 | PMSR2 | Aliphatic Glucosinolate | Genetic |
| AT5G07470 | PMSR3 | Aliphatic Glucosinolate | Genetic |
| AT1G07640 | DOF1.1 | Aliphatic Glucosinolate | Genetic |
| AT3G09710 | IQD1 | Aliphatic Glucosinolate | Genetic |
| AT5G07690 | MYB29 | Aliphatic Glucosinolate | Genetic |
| AT5G07700 | MYB76 | Aliphatic Glucosinolate | Genetic |
| AT5G61420 | MYB28 | Aliphatic Glucosinolate | Genetic |
| AT5G61640 | PMSR1 | Aliphatic Glucosinolate | Genetic |
| AT4G39950 | CYP79B2 | Indolic Glucosinolate | Biochem |
| AT2G22330 | CYP79B3 | Indolic Glucosinolate | Biochem |
| AT1G74100 | ATST5A | Indolic Glucosinolate | Biochem |
| AT1G74090 | ATST5B | Indolic Glucosinolate | Biochem |
| AT5G60890 | ATR1/MYB34 | Indolic Glucosinolate | Genetic |

| | | | |
|---|---|---|---|
| AT5G46760 | ATR2/bHLH | Indolic Glucosinolate | Genetic |
| AT2G30870 | AtGSTF10 | Indolic Glucosinolate | Homology |
| AT2G30860 | AtGSTF07 | Indolic Glucosinolate | Homology |
| AT4G31500 | CYP83B1 | Indolic Glucosinolate | Biochem |
| AT1G18570 | MYB51 | Indolic Glucosinolate | Genetic |
| AT5G57220 | CYP81F2 | Indolic Glucosinolate | Biochem |
| AT4G37400 | CYP81F3 | Indolic Glucosinolate | Homology |
| AT4G37410 | CYP81F4 | Indolic Glucosinolate | Homology |
| AT4G37430 | CYP81F1 | Indolic Glucosinolate | Homology |
| AT3G26830 | PAD3/CYP71B15 | Camalexin | Biochem |
| AT2G30770 | CYP71A13 | Camalexin | Biochem |
| AT2G30750 | CYP71A12 | Camalexin | Homology |
| AT2G45570 | CYP76C2 | Camalexin | Guess |
| AT1G11610 | CYP71A18 | Camalexin | Homology |
| AT1G58260 | CYP79C3 | Unknown GLS | Homology |
| AT1G58265 | CYP79C2 | Unknown GLS | Homology |
| AT1G79370 | CYP79C1 | Unknown GLS | Homology |
| AT5G35917 | CYP79A3 | Unknown GLS | Homology |
| AT5G35920 | CYP79A4 | Unknown GLS | Homology |
| AT1G07780 | PAI | Tryptophan | Guess |
| AT1G25220 | ASB | Tryptophan | Guess |
| AT1G29410 | PAI | Tryptophan | Guess |
| AT2G04400 | I3GPS | Tryptophan | Guess |
| AT2G29690 | ASA2 | Tryptophan | Guess |
| AT3G54640 | TSA2 | Tryptophan | Guess |
| AT4G02610 | TSA | Tryptophan | Guess |
| AT4G27070 | TSB | Tryptophan | Guess |
| AT5G05730 | ASA1 | Tryptophan | Guess |
| AT5G17990 | PAT | Tryptophan | Guess |
| AT5G48220 | I3GPS | Tryptophan | Guess |
| AT5G05590 | PAI | Tryptophan | Guess |
| AT1G08700 | homoserine kinase | methionine biosynthesis | Guess |
| AT2G17265 | homoserine kinase | methionine biosynthesis | Guess |
| AT3G03780 | methionine synthase | methionine biosynthesis | Guess |
| AT3G22740 | homocysteine S-methyltransferase | methionine biosynthesis | Guess |
| AT3G25900 | homocysteine S-methyltransferase | methionine biosynthesis | Guess |
| AT3G63250 | homocysteine S-methyltransferase | methionine biosynthesis | Guess |
| AT4G11610 | homoserine kinase | methionine biosynthesis | Guess |
| AT5G20980 | | methionine biosynthesis | Guess |
| AT1G02500 | methionine adenosyltransferase | methionine degradation I | Guess |
| AT2G36880 | methionine adenosyltransferase | methionine degradation I | Guess |
| AT3G17390 | methionine adenosyltransferase | methionine degradation I | Guess |
| AT3G23810 | adenosylhomocysteinase | methionine degradation I | Guess |
| AT4G01850 | methionine adenosyltransferase | methionine degradation I | Guess |
| AT4G13940 | | methionine degradation I | Guess |
| AT1G19920 | ATP sulfurylase | sulfate assimilation | Guess |
| AT1G62180 | APS reductase | sulfate assimilation | Guess |
| AT3G22890 | ATP sulfurylase | sulfate assimilation | Guess |
| AT4G04610 | APS reductase | sulfate assimilation | Guess |
| AT4G14680 | ATP sulfurylase | sulfate assimilation | Guess |

| | | | |
|---|---|---|---|
| AT4G21990 | APS reductase | sulfate assimilation | Guess |
| AT5G04590 | sulfite reductase | sulfate assimilation | Guess |
| AT5G43780 | ATP sulfurylase | sulfate assimilation | Guess |
| AT2G14750 | AKN1 | sulfate assimilation | Guess |
| AT3G03900 | AKIN3 | sulfate assimilation | Guess |
| AT5G67520 | AKIN2 | sulfate assimilation | Guess |
| AT1G22410 | 3-deoxy-7-phosphoheptulonate synthase | HomoCys | Guess |
| AT1G33320 | cystathionine gamma-synthase | HomoCys | Guess |
| AT1G55880 | | HomoCys | Guess |
| AT1G64660 | | HomoCys | Guess |
| AT3G01120 | cystathionine gamma-synthase | HomoCys | Guess |
| AT3G10050 | cystathionine beta-synthase | HomoCys | Guess |
| AT3G22460 | cystathionine beta-synthase | HomoCys | Guess |
| AT3G57050 | cystathionine beta-lyase | HomoCys | Guess |
| AT4G23600 | cystathionine beta-lyase | HomoCys | Guess |
| AT5G28020 | | HomoCys | Guess |
| AT5G28030 | | HomoCys | Guess |
| AT1G55920 | serine acetyltransferase | cysteine biosynthesis | Guess |
| AT2G17640 | serine O-acetyltransferase | cysteine biosynthesis | Guess |
| AT2G34970 | serine O-acetyltransferase | cysteine biosynthesis | Guess |
| AT2G43750 | O-acetylserine (thiol) lyase | cysteine biosynthesis | Guess |
| AT3G03630 | cysteine synthase | cysteine biosynthesis | Guess |
| AT3G04940 | cysteine synthase | cysteine biosynthesis | Guess |
| AT3G13110 | serine acetyltransferase | cysteine biosynthesis | Guess |
| AT3G59760 | O-acetylserine (thiol) lyase | cysteine biosynthesis | Guess |
| AT3G61440 | cysteine synthase | cysteine biosynthesis | Guess |
| AT4G14880 | O-acetylserine (thiol) lyase | cysteine biosynthesis | Guess |
| AT4G29540 | serine O-acetyltransferase | cysteine biosynthesis | Guess |
| AT5G38530 | cysteine synthase | cysteine biosynthesis | Guess |
| AT5G56760 | serine acetyltransferase | cysteine biosynthesis | Guess |
| AT4G23100 | glutamate-cysteine ligase | Glutathione Synthesis | Guess |
| AT5G27380 | glutathione synthetase | Glutathione Synthesis | Guess |

Table 2.14: Glucosinolates which can be produced from inorganic nutrients in the model. Representative glucosinolates from each of the three classes can be produced, and various sizes of elongated aliphatic glucosinolates. Genes which affect the predicted ability to produce any of these glusosinolates was reported.

| Modelled glucosinolates |
|---|
| **Aliphatic (from methionine)** |
| ***homomethionine*** |
| 3-methylthiopropyl-desulfo-glucosinolate |
| 3-methylthiopropyl-glucosinolate |
| 3-methylsulfinylpropyl-glucosinolate |
| 3-hydroxypropyl-glucosinolate |
| 3-benzoyloxypropyl-glucosinolate |
| ***dihomomethionine*** |
| 4-methylthiobutyldesulfoglucosinolate |
| 4-methylthiobutyl glucosinolate |
| 4-methylsulfinylbutyl glucosinolate |
| ***trihomomethionine*** |
| 5-methylthiopentylglucosinolate |
| 5-methylsulfinylpentyl glucosinolate |
| 4-pentenylglucosinolate |
| **emphtetrahomomethionine** |
| 6-methylthiohexyldesulfoglucosinolate |
| 6-methylthiohexylglucosinolate |
| ***pentahomomethionine*** |
| 7-methylthioheptyldesulfoglucosinolate |
| 7-methylthioheptyl glucosinolate |
| 7-methylsulfinylheptyl glucosinolate |
| ***hexhomomethionine*** |
| 8-methylthiooctyldesulfoglucosinolate |
| 8-methylthiooctyl glucosinolate |
| 8-methylsulfinyloctyl glucosinolate |
| |
| **Indolic (from Tryptophan)** |
| indolylmethyl-glucosinolate |
| indolylmethyl glucosinolate aglycone |
| 4-methoxy-3-indolylmethyl-glucosinolate |
| 4-methoxy-3-indolylmethyl glucosinolate aglycone |
| 4-hydroxy-3-indolylmethyl-glucosinolate |
| |
| **Aromatic (from phenylalanine)** |
| benzyl-desulfoglucosinolate |
| glucotropeolin |

Table 2.16: The genes which were expected to affect glucosinolate phenotypes, and are included in the model, but are not predicted by FBA using the 'gene knockout, no biomass requirement' approach. A large number of the genes are involved in degradation which is not well captured by flux balance analysis. The other failed predictions are likely due to errors in the model reaction structure, or incomplete mapping of the identified gene to the reactions which it catalyses.

| AGI | Gene name | Pathway | Evidence |
|---|---|---|---|
| AT4G03060 | AOP2 | Aliphatic Glucosinolate | Biochem |
| AT3G03190 | AtGSTF11 | Aliphatic Glucosinolate | Guess |
| AT1G65880 | BZO1 | Aliphatic Glucosinolate | Biochem |
| AT2G30870 | AtGSTF10 | Indolic Glucosinolate | Homology |
| AT2G22330 | CYP79B3 | Indolic Glucosinolate | Biochem |
| AT2G30860 | AtGSTF07 | Indolic Glucosinolate | Homology |
| AT4G39950 | CYP79B2 | Indolic Glucosinolate | Biochem |
| AT5G35917 | CYP79A3 | Unknown GLS | Homology |
| AT1G58265 | CYP79C2 | Unknown GLS | Homology |
| AT5G35920 | CYP79A4 | Unknown GLS | Homology |
| AT2G30750 | CYP71A12 | Camalexin | Homology |
| AT2G30770 | CYP71A13 | Camalexin | Biochem |
| AT1G11610 | CYP71A18 | Camalexin | Homology |
| AT3G26830 | PAD3/CYP71B15 | Camalexin | Biochem |
| AT4G29540 | serine O-acetyltransferase | cysteine biosynthesis | |
| AT3G61440 | cysteine synthase | cysteine biosynthesis | |
| AT3G10050 | cystathionine beta-synthase | HomoCys | |
| AT1G64660 | cystathionine beta-lyase | HomoCys | |
| AT3G25900 | homocysteine S-methyltransferase | methionine biosynthesis | |
| AT3G03780 | methionine synthase | methionine biosynthesis | |
| AT3G22740 | homocysteine S-methyltransferase | methionine biosynthesis | |
| AT3G63250 | homocysteine S-methyltransferase | methionine biosynthesis | |
| AT4G13940 | adenosylhomocysteinase | methionine degradation I | |
| AT4G01850 | methionine adenosyltransferase | methionine degradation I | |
| AT3G23810 | adenosylhomocysteinase | methionine degradation I | |
| AT1G02500 | methionine adenosyltransferase | methionine degradation I | |
| AT3G17390 | methionine adenosyltransferase | methionine degradation I | |
| AT2G36880 | methionine adenosyltransferase | methionine degradation I | |
| AT1G51470 | TGG5 | GLS Breakdown | Biochem |
| AT5G25980 | TGG2 | GLS Breakdown | Biochem |
| AT1G54000 | myrosinase-associated protein | GLS Breakdown | Homology |
| AT1G54030 | GDSL-motif lipase family protein | GLS Breakdown | Homology |
| AT1G66280 | glycosyl hydrolase family 1 protein | GLS Breakdown | Homology |
| AT1G47600 | TGG4 | GLS Breakdown | Biochem |
| AT1G54010 | AGG1/ESM2 | GLS Breakdown | Biochem |
| AT5G26000 | TGG1 | GLS Breakdown | Biochem |
| AT5G48375 | TGG3 | GLS Breakdown | Biochem |
| AT3G09260 | PYK10 | GLS Breakdown | Biochem |
| AT1G52400 | BGL1 | GLS Breakdown | Homology |
| AT1G51490 | TGG6 | GLS Breakdown | Biochem |

# Chapter 3

# The use of elementary modes for analysis of nutritional requirements

The calculation of elementary flux modes (EFMs) provides a mathematical framework for the analysis of metabolic models. Unlike flux balance analysis, elementary modes do not impose an artificial objective upon the metabolic system, and can therefore be used to analyse the full metabolic space imposed by the stoichiometry of the model. However until very recently, calculation of elementary modes was not feasible for genome scale models.

Here we consider sets of elementary modes calculable for the genome scale model of *Arabidopsis thaliana* presented in chapter 2. We demonstrate that although it is still not computationally feasible to calculate all modes, the calculable subsets approximate the behaviour of the full set, and apply them to the study of reaction relatedness, nutrient use efficiency, and nutrient requirement tradeoffs. We find evidence for surprisingly little nutrient use flexibility through metabolic regulation, but that access to external resources in terms of nitrogen and energy sources apparently dictate metabolic flexibility.

## 3.1 Calculable EFM subsets can be used to approximate the behaviour of the full set

The number of EFMs in a given network depends on the structure of the network, however it is generally understood that the number of EFMs undergoes a 'combinatorial explosion' with the number of reactions. It is, therefore, expected that genome scale models of metabolism contain billions of EFMs. This

has proven to be a major hurdle to the widespread application of EFM analysis, and we are not aware of any study in which they have been used for the analysis of a genome scale model of plant metabolism.

The recently published, TreeEFM algorithm [163], (see Methods) represents a significant advance in the efficient computation of elementary modes. Here it was used to calculate EFMs of the Arabidopsis model presented in chapter 2. However, Figure 3.1 demonstrates that the complete enumeration of all elementary modes is still not practically possible, even with the TreeEFM software. This is without consideration of the further difficulty of storage and analysis which would arise were all EFMs calculable. As well as taking quite a large amount of time to return only a small fraction of EFMs, Figure 3.1 shows that the relationship between time and number of EFMs found is non-linear. This is likely a consequence of the increasingly constrained linear models becoming harder to solve (see Methods), and indicates that the complete EFM set cannot be practically calculated by simply running the program for longer.

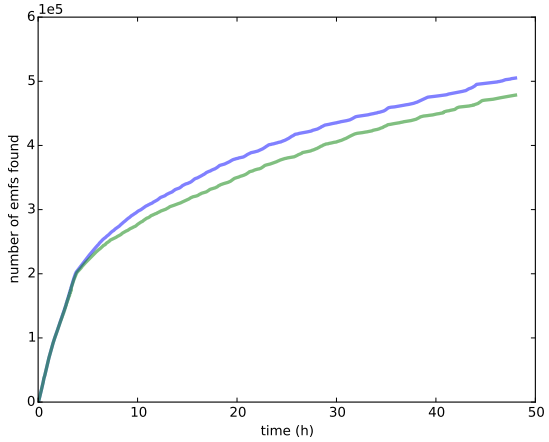



Figure 3.1: It is not currently practically possible to compute all EFMs in the Arabidopsis model. The number of EFMs which produce biomass, recovered in two independent runs of TreeEFM, from the genome scale *Arabidopsis thaliana* consisting of 3,124 reactions, presented in chapter 2. It takes several days to calculate tens of thousands of modes, whereas it is expected that billions of modes may exist. Furthermore, the time to calculate each mode increases with the number that have previously been found.

In order to apply EFM analysis to the Arabidopsis model, we must therefore establish that the properties of the full EFM set can be approximated by a calculable subset. In order to achieve this, we wished to establish that the features of the subset are stable with respect to size. Metrics perviously used to asses the quality of calculated EFM subsets are the length distribution of the set, and the fraction of found EFMs each reaction participates in [47, 131].

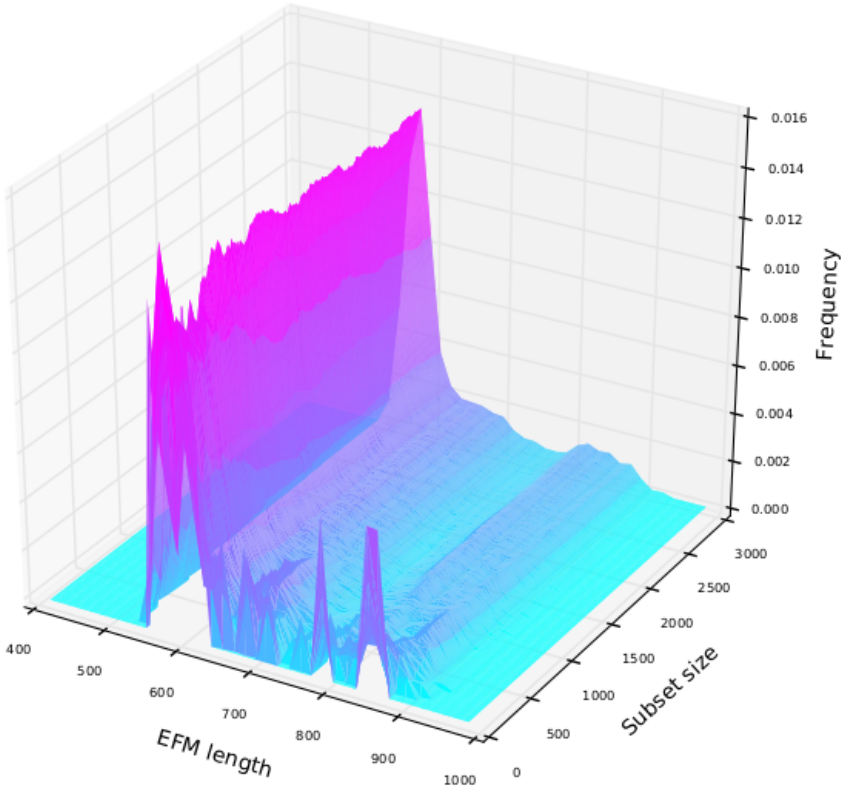



Figure 3.2: The length distribution of subsets of EFMs are stable to changes in subset size, indicating that the length distribution in calculable subsets approximate the behaviour of the full set.

Figure 3.2 shows that the length distribution of EFMs stabilises extremely rapidly as subset size increases. After approximately 500 EFMs are calculated, the length distribution does not change significantly over all subset sizes which we calculated (up to 300,000 EFMs), and can therefore be expected to be the same as in the full set.

Figure 3.3a shows the reaction participation fraction in EFMs for all reactions as EFM subset size increases. For a given reaction this is calculated as the fraction of EFMs in the subset in which that reaction carries non-zero flux. Reaction participation can be seen to be fairly dynamic, in terms of subset size, for subsets of less than approximately 10,000 EFMs, but is moderately stable above approximately 120,000 EFMs, as the majority of lines become much straighter. It should be noted that even within this more stable region, reaction participation for some reactions continue to change, and can switch suddenly

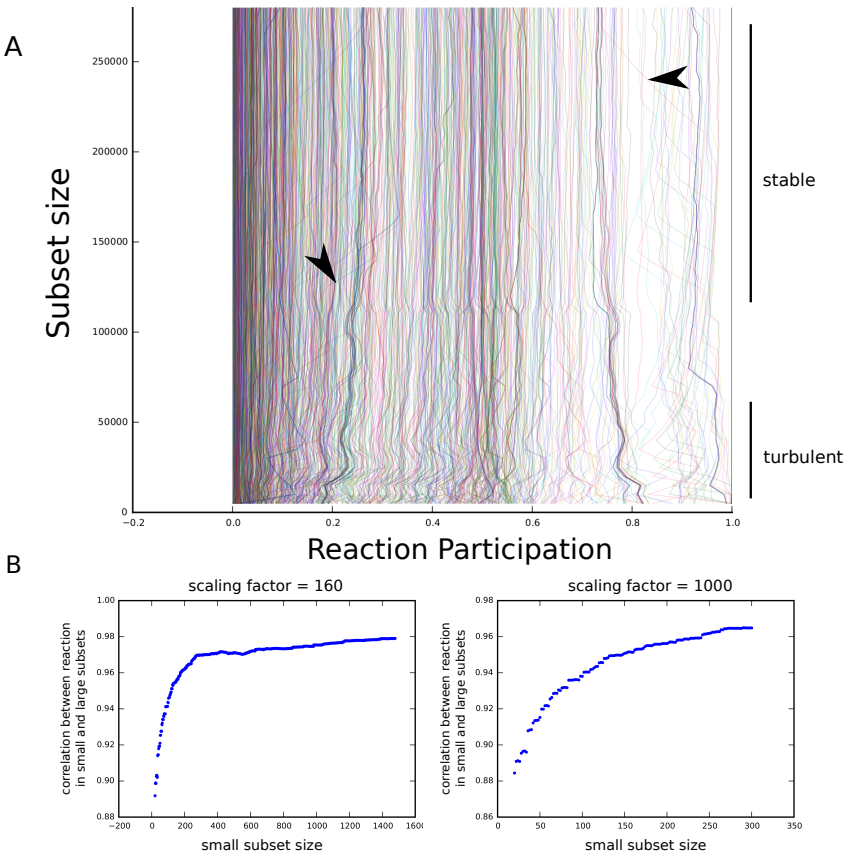



Figure 3.3: The reaction participation fractions in subsets of EFMs are stable with subset size, indicating that calculable subsets approximate the behaviour of the full set. A, the fraction of the found EFMs in which each reaction participates (carries non-zero flux). Each line corresponds to a single reaction. For EFM subsets of more than approximately 100,000 modes, reaction participation is mostly consistent across subset sizes, although switching behaviour continues to occur. Arrows indicate examples of reactions which switch from apparently stable, to dynamic reaction participation behaviour as subset size increases. B, correlation in reaction participation between small and large EFM subsets is high. Each point corresponds to the Pearson's correlation coefficient between reaction participation in a small EFM subset and large EFM subset. The large subset consists of 160 (left), or 1,000 (right) times the number of EFMs in the small subset. Correlation is greater, the smaller the scaling factor, and rapidly saturates with subset size.

from apparently stable, to dynamic behaviour (examples of this behaviour are highlighted by arrows). This is due to the tree exploration strategy of TreeEFM, in which solution branches constrained to involve underrepresented reactions in previous solutions are prioritised for exploration (see Methods), leading to sudden switches in reaction participation.

Although Figure 3.3a clearly shows that reaction participation stabilises over the calculated subset sizes, the degree of this stabilisation is difficult to interpret. Figure 3.3b, plots correlation between reaction participation in relatively small, and large subsets, with the scaling factor between the number of EFMs per subset indicated. It can be seen that even though the small subsets considered are within the 'turbulent' region identified in Figure 3.3a, the correlation coefficient to the larger set is $>0.9$ for all sets of EFMs larger than 50. Although correlation is greater the smaller the scaling factor, this extremely strong relationship holds even between subsets differing in size 1000-fold. Furthermore, correlation increases with the size of the smaller subset, suggesting that the calculable 300,000 EFM subsets used for analysis in this chapter are extremely representative of the full EFM set of the model. These coefficients compare favourably with previously published values of a 'representative' EFM subset [131].

Figure 3.3b also shows that the increase in correlation with subset size saturates, suggesting that enumeration of EFMs offers diminishing returns in terms of new information, although as already discussed in relation to Figure 3.1 they take increasing time to calculate. Therefore although Figure 3.3a indicates that a small number of reactions exhibit the concerning switching behaviour, these are unlikely to significantly affect any analysis.

We have shown that subsets of EFMs calculated using TreeEFM well represent larger sets in terms of two metrics commonly used to evaluate EFM subset quality. However, large section of analysis presented later in this chapter is concerned with correlation between reaction flux across the calculated EFMs. In Figure 3.4, we therefore consider how the Pearson's correlation coefficient between all pairwise combinations of reactions relates between smaller and larger subsets. It can be seen that even the first 10,000 elementary modes calculated are very representative of a much larger set, as indicated by the strong diagonal line in the right hand facet, and that again, as the number of EFMs considered increases, the relationship becomes stronger. It is therefore expected that the full 300,000 EFM subset used is even more strongly related to a similarly larger set.

We cannot compare the quality of the calculable subsets to the full, true set of EFMs, however, the stability of the considered metrics with respect to increasing subset size, and correlation between reaction correlation coefficients in small and large subsets indicates that the subsets returned by TreeEFM approximate the behaviour of the full set, and that therefore the calculated EFM subsets can reasonably used to analyse the behaviour of the system. We acknowledge that small subsets calculated using the TreeEFM approach may be more similar to

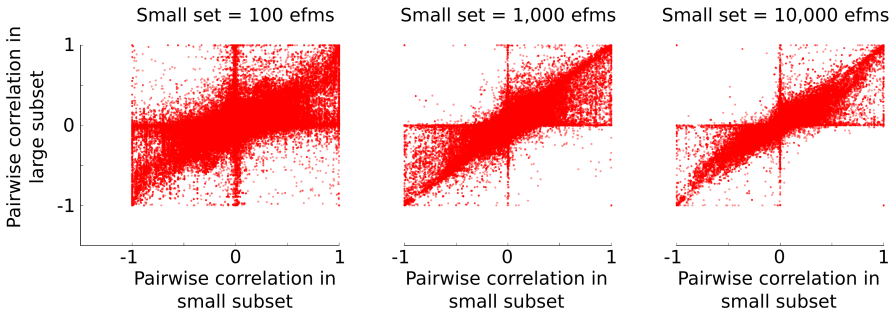



Figure 3.4: Correlation coefficients between reaction pairs in small EFM subsets are strongly indicative of correlation between reaction pairs in large subsets. Each point is the Pearson's correlation coefficient between flux in a pair of reactions across the considered elementary modes. In each graph the large subset consists of 10 times as many elementary modes as the small subset. Small subsets of different sizes are shown, and the strength of the relation between the small and large set increases with subset size. This suggests that correlation coefficients between pairs of reactions derived from the calculable EFM subsets can be used to approximate the correlation coefficients across the full set of EFMs.

larger sets calculated using the TreeEFM approach, than an equivalent sized set calculated using a different method, due to the internal prioritisation of EFMs for calculation within the TreeEFM algorithm. However, in our hands the application of other methods to calculate large numbers of EFMs using the Arabidopsis model was found to be impractically demanding of computational resources. In the rest of this chapter, subsets of 300,000 EFMs are analysed as a compromise between computation time and fidelity of approximation to the full set.

## 3.2 Reaction correlation analysis

### 3.2.1 Introduction

The potential usefulness of EFMs for metabolic flux analysis is well established [189]. However, as network size, and consequently the number of EFMs increases, they become increasingly difficult not only to calculate, but also to interpret. These difficulties mean that although EFMs have been used successfully for a number of applications in plants [193, 178, 167, 18], we are not aware of any studies which have considered more than a small subset of reactions in their model. Having seen that newly developed methods allow the calculation of sufficiently large subsets of elementary modes to approximate the behaviour of the full set, we now turned to their interpretation. The output of the analysis

consisted of 300,000 elementary modes, each consisting of flux through more than 3,000 reactions, and is therefore extremely complicated to interpret. Here we applied unsupervised statistical learning methods to facilitate the analysis of this dataset.

One common approach to the simplification of complex systems is modularisation, in which by grouping similar behaving elements together we can reduce the number of parts considered. This suggests two approaches to simplify the interpretation of EFMs. Firstly, we can group similar EFMs themselves, based upon the similarity of participating reaction sets in order to reduce the number of EFMs which must be considered. This is equivalent to the clustering of similar minimal t-invariants described by Grafahrend-Belau et al. [72]. However while this has proven somewhat useful in grouping components of regulatory cascades together, often elementary flux modes of metabolic networks do not cluster well due to the presence of multiple essentially independent reaction motifs [160], caused by, for example independent, parallel reaction pathways, and compartmentalisation.

This suggests that the second approach, in which we analyse reaction relatedness, and group similar reactions into co-occuring motifs, may be more useful. This potentially allows an EFM to be more simply analysed in terms of the presence or absence of these motifs, the number of which are strictly less than the number of reactions, but additionally, and what we find more appealing, is that 'functions' can potentially be ascribed to these groups, simplifying their interpretation.

This reaction grouping is similar to the approach of grouping reactions into functionally related 'metabolic pathways' which has historically been important for the interpreting metabolic flux, and the response of the system to perturbation. At one level, manually derived metabolic pathways can be considered a 'gold standard', in terms of their accuracy, due to the direct experimental evidence which has led to the association of their member reactions. However cellular metabolism is often much more plastic than textbook representations of metabolic pathways [189], and classical metabolic pathways are not a perfect representation of metabolic modes. Previous EFM studies have already demonstrated that alternative functions and reaction groupings can be conceptually useful [167].

Various algorithmic approaches to modularisation have been proposed, broadly based on: direct analysis of network topology, flux coupling estimated through flux balance analysis, or through examination of elementary modes (see [101] for review). In terms of EFMs, reactions have previously been grouped according to their exact co-occurence [181], i.e. that the presence of one reaction in an elementary mode is necessary and sufficient for the presence of another. This results in generation of the smallest "biologically meaningful" motifs. However, this approach does not capture the inherently hierarchical nature of reaction relatedness, (for example, the concept of 'pathways' versus 'superpathways'), and is of limited use for simplifying large networks due to the extremely small mo-

tifs produced. Peres et al. [160] propose an alternative motif finding algorithm which allows more flexibility in specifying the allowed relatedness of grouped reactions, but in our hands this algorithm is too slow to be practical for the identification of motifs in large numbers of EFMs consisting of large numbers of reactions.

Here, we use a motif finding method more similar to the kinds of clustering approaches traditionally used in the analysis of gene expression profiles [11], in which transcripts are hierarchically clustered by the correlation coefficient of their expression across experiments. In our approach, this translates to the correlation coefficient of predicted flux between reactions across the calculated EFMs. This is related to flux balance analysis based methods [164, 238, 32], in which coupling between reactions is estimated within the flux space permitted under some optimality criterion, except that an EFM based approach more fully considers the full flux space allowed under the metabolic steady state constraint.

This is also similar to the pioneering null-space based approach of Poolman et al. [170]. To briefly refresh topics discussed in chapter 1; the null-space of a matrix $S$, is the set of all vectors, $v$, such that $S \cdot v = 0$. In the context of flux analysis, where $S$ is the stoichiometric matrix, the null-space is the set of all steady state flux solutions. All feasible steady states, including elementary modes are included in the null-space, and in geometric terms EFMs are the extremal vectors of the null-space [237]. Poolman et al. [170] propose a method in which a basis of the null-space kernel of the stoichiometric matrix is analysed in order to derive the correlation coefficients between reactions across all elementary modes, and then use these coefficients to cluster reactions in several genome scale models of bacteria.

Our approach differs in that, whereas Poolman et al. consider correlated expression across all elementary modes, we consider only those which are able to generate biomass, as indicated by flux through the biomass equation. Our approach therefore occupies an intermediate space between the optimality assumptions imposed by flux balance analysis, which may restrict the considered metabolic modes too severely, and the unconstrained nature of the solution of the Poolman method, which is likely to be adversely affected by EFMs which are not biologically utilised, and generate spurious reaction relationships, and hide more realistic ones.

### 3.2.2 Analysis of correlation coefficient distributions

Before using correlation coefficients to group related reactions together, we initially consider the distribution of correlation coefficients between reactions. This preliminary analysis does not allow detailed conclusions in terms of particular reactions, or flux distributions. However, it does allow a comparison of the consequences of network structure between species on the potential independence of

reactions. It also demonstrates differences between results generated using our method, and the Poolman, null-space method [170] justifying the calculation of EFM sets. Furthermore, it also shows that the reaction relationship structure suggested by traditional biochemical pathways fails to reflect experimental relationship data.

#### 3.2.2.1 Comparison to null-space analysis of bacteria

A potentially interesting comparison to make, is between the distribution of correlation coefficients in models of bacterial species and Arabidopsis. This allows an assessment of the extent to which reaction co-expression is imposed in different kingdoms as a consequence reaction network topology, and therefore of metabolic flexibility expressed through reaction independence. Poolman et al. [170], consider two bacterial species, *Escherichia coli*, and *Streptomyces coelicolor*. As reproduced in Figure 3.5b, absolute correlation coefficients in both of these models were seen to follow a log-normal distribution. To enable direct comparison to this result, we here use the null-space approach as well.

Figure 3.5a shows that relative to bacterial systems, in Arabidopsis there are proportionally fewer highly correlated reactions, and a long tail of weakly related reaction pairs. This suggests that relative to microbial species, the Arabidopsis reaction network allows more flexibility in reaction co-expression. This is likely to be at least partially a consequence of the subcellular compartmentalisation within the Arabidopsis model. Compartmentalisation leads to a larger model in terms of reaction number (thousands rather than hundreds of reactions in the considered models), but additionally, copies of reactions in different compartments can act in a mutually compensatory manner, which causes a reduced correlation between them, and related reactions.

Hosseini et al. [90] note that flux coupling in *E. coli* apparently allows somewhat modular control of metabolism by regulation of key 'root' fluxes, which in turn control many subsidiary reactions indirectly, through network topology constraints. The relative scarcity of highly correlated reaction sets in Arabidopsis suggests, perhaps unsurprisingly that more sophisticated regulatory controls are possible, but also required in higher organisms, as reactions are generally more independent, thus allowing higher resolution regulation of metabolism, and greater potential metabolic flexibility.

#### 3.2.2.2 Comparison of null-space derived, and directly calculated correlation coefficients

The null-space method proposed by Poolman et al. [170] is able to calculate reaction correlation coefficients more rapidly than calculating them directly from elementary modes. In order to determine whether the extra steps of calculating

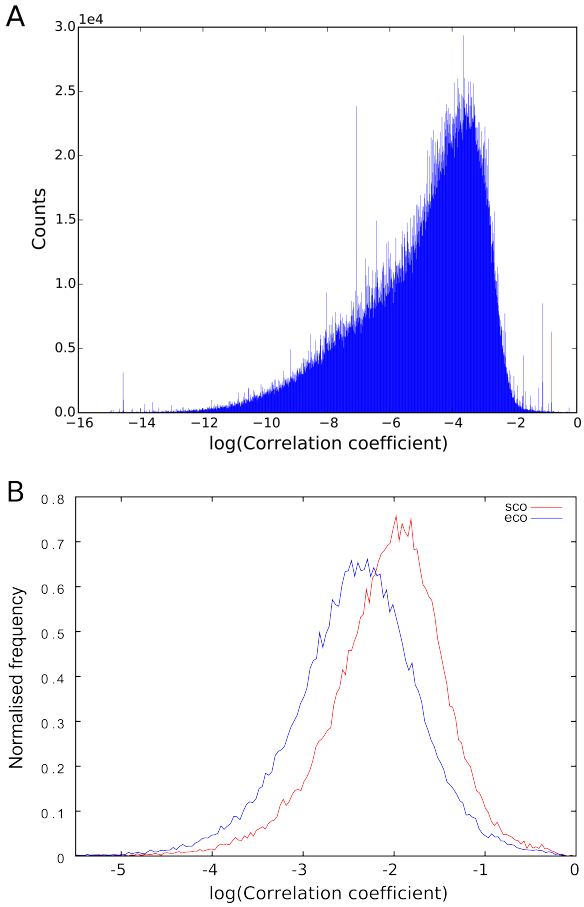



Figure 3.5: Distribution of correlation coefficients calculated using the null-space method. A, The distribution of log transformed absolute correlation coefficients across all reactions in the Arabidopsis. B, The distribution of log transformed, absolute correlation coefficients in *Escheria coli* (blue), and *Streptomyces coelicolor* (red). B is reproduced from [170]. Comparison of A and B indicates that the Arabidopsis model exhibits a large tail of weakly related reactions relative to the bacteria.

correlation coefficients from biomass producing EFMs is worthwhile, we compared the reaction correlation coefficients calculated using the null-space method [170] to those calculated from biomass producing EFMs in order to see whether the different approaches leads to different expected reaction relationships.

Figure 3.6 shows a lack of agreement between correlation coefficients between reaction pairs calculated using the null-space method [170], and directly from flux EFMs. Mathematically, these metrics are equivalent when considering all EFMs [170], and therefore the disagreement arises because we are considering a subset of elementary modes. This disagreement is either because 1.) the calculated subsets of elementary modes are not sufficient to represent the behaviour or the full set, or 2.) because we only calculate elementary modes which involve the biomass reaction, rather than all EFMs. We have already seen in section 3.1 evidence that the properties of the EFM subset are stable with regards to the number of EFMs calculated , and that therefore scenario 1 is unlikely to be the cause of the observed disagreement. We therefore consider option 2 as likely to be the dominant factor.

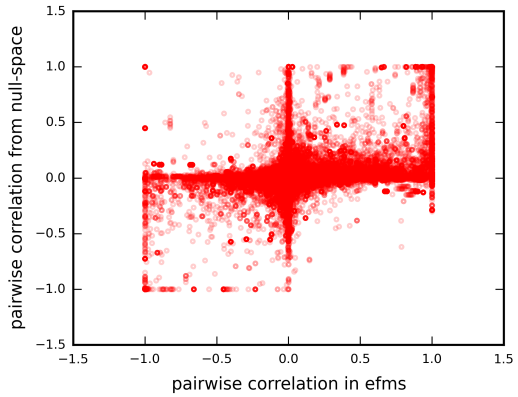



Figure 3.6: The dissimilarity between reaction correlation coefficients calculated using the null-space method, and directly from EFMs. Each point corresponds Pearson's correlation between a reaction pair. This suggests that our method, in which correlation coefficients between reactions are calculated only for EFMs which produce biomass, produces significantly different results to the method of Poolman et al.[170] which uses null-space analysis to consider correlation between reaction fluxes in all elementary modes.

Figure 3.6 also shows that the null-space based approach tends to derive a greater proportion of weakly correlated reactions than is seen using direct calculation from biomass producing EFMs, as indicated by the relative number of points proximal to the horizontal, as opposed to vertical axis. This makes sense, as the null-space approach considers, for example, futile reaction cycles comprising small sets of reactions, in which the vast majority of reactions carry no flux, and other small, independent, biologically implausible modes. These implausible modes are likely to correspond to a significant proportion of the

elementary modes considered [64], and obscure reaction relationships in more biologically likely elementary modes.

In comparison the constraint that EFMs must produce biomass imposes more structure on the flux modes through the system, and is less likely to obscure reaction relatedness in this way. Identifying biologically likely subsets of elementary modes is not trivial, and we do not suggest that the calculated set of biomass producing modes are all biologically plausible. However, they are likely to be more representative on average than the full set of all EFMs. Indeed, Poolman et al. [170] speculate that the relative lack of modularity in the models they studied could be due to studying all EFMs, including biologically unlikely ones.

We have shown that there is a difference in results generated using the different methods, and seen some advantage to calculating correlation through elementary modes directly, rather than through null-space analysis. It is therefore interesting to move forward, and consider reaction clustering using the correlation coefficients calculated from EFMs.

#### 3.2.2.3 The distribution of EFM derived reaction relatedness measures is more similar to experimental data than metabolic pathway databases

Several manually curated databases exist which group reactions into particular groups, reflecting conventional views of metabolic pathways. In order to initially evaluate whether the EFM derived reaction relatedness measures provide any information not captured in these databases, we compared the distribution of relatedness metrics derived from these databases, and from EFMs, to correlation coefficients across transcriptomic experiments. We use transcriptomic data as a proxy for flux through the enzyme catalysed reaction. Although transcriptomic data is not an ideal measure of metabolic reaction flux, its abundant availability, and the success of its use in a number of network inference [166], and flux balance [220] studies justifies its use.

Figure 3.7 shows the distribution of reaction pair distances derived from two metabolic pathway databases: Aracyc [253], and KEGG [50], from transcriptomic data, (from the Expression Atlas database [161]), and from EFMs, (see Methods for details of the reaction pair distance metrics used). The distributions derived from manually curated reaction databases, (Figure 3.7a,b) are strikingly dissimilar to the 'true' relatedness distribution, derived from transcriptomic expression data (Figure 3.7c).

As shown in Figure 3.7a, in the Aracyc database [253], the modal distances are 1.0, i.e. completely unrelated, followed by 0.0, i.e. completely identical. Other, intermediate relatedness distances are uncommon. Figure 3.7b shows that similar pattern is observed in the KEGG database; although a larger number of intermediate distances is observed, the vast majority of reactions are

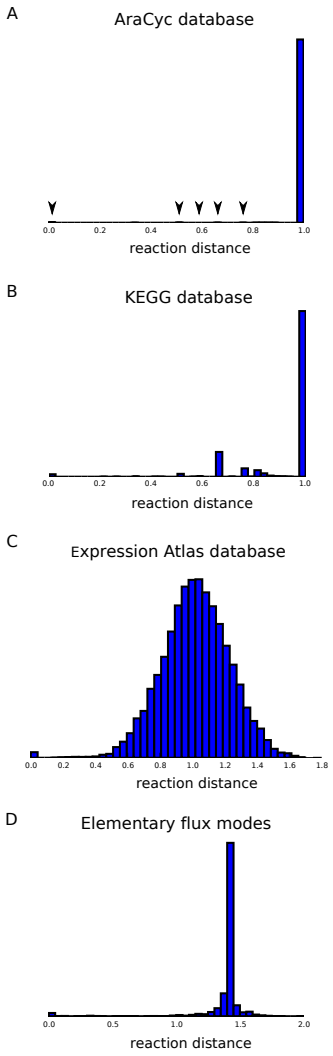



Figure 3.7: Pairwise reaction distance distributions derived from transcriptomic data are more similar to those calculated from EFMs, than from manually curated databases. The distance distributions between all pairwise combinations of reactions in A, the Aracyc database, B, KEGG database, C, Expression Atlas transcriptomic database, D, the first 300,000 recovered EFMs. See Methods for details of the databases, and reaction distance metrics used. Uncommon, non-zero reaction distances are indicated by arrows in A. Transcriptomic and EFM derived distances have a unimodal distribution, centred upon an intermediate distance indicating that most reactions are somewhat related to most others. Conversely, the modal value in both manually curated databases is 1.0, indicating that reactions are completely unrelated. This suggests that EFM derived distances might capture information about the biological system that is missing in the manually curated databases.

completely unrelated, with a distance of 1.0. Conversely, in the Expression Atlas [161] transcriptomic dataset, (Figure 3.7c), pairwise transcript expression distances follow a normal distribution, centred about halfway between complete relatedness and unrelatedness. Under the assumption that transcript abundance is related to flux through the reaction(s) associated with the gene product, this indicates: firstly, that flux through almost all reactions is somewhat predictive of flux through most others (as indicated by the generally intermediate distance values), and secondly, that reactions have a much more contiguous, hierarchical degree of relatedness, than is captured in the databases, as indicated by the spread of different relatedness values.

The (completely correlated) peak at 0.0 distance in Figure 3.7c is an artefact of the mapping between transcript and reaction identities, in which the same transcript species is mapped to all instances of the same reaction in multiple compartments when no specific compartmental information is available. Generally, reactions are much less strongly predictive of the most related reactions than is suggested in the databases. This further indicates the crudity of reaction relatedness information identified in 'classical' biological pathways.

The transcriptomic distribution is based on somewhat noisy experimental data, and so the normal distribution observed could be partly due to this noise causing a spread around much more discrete relatedness levels. However the difference in the modal pairwise reaction distance clearly shows that the relatedness structure is different between the reaction-path databases and transcriptomic data.

Conversely, the EFM derived distance distribution (Figure 3.7d), is qualitatively similar to transcript expression distribution, in that it has an approximately normal distribution, centred at an intermediate relatedness value. Therefore the EFM derived reaction relatedness measures are indeed correctly capturing some information not in the reaction databases, and the conventional reaction grouping of metabolic paths.

The elementary mode distribution is clearly not identical to the transcriptomic one. The distributions are centred at different points, and the standard deviation is greater in the experimental data, indicating that the spread of partial, hierarchical relatedness seen is not as great in the EFM derived metric. Although it is difficult to identify the contributions of the various causes of this discrepancy, it is likely due to a combination of model and mapping inaccuracies, noisy experimental data, and the disconnect between transcript abundance and flux through the enzyme catalysed reaction. Nevertheless the distribution of relatedness in Figure 3.7 is clearly more similar between transcript and EFMs, than between transcript and path databases.

### 3.2.3 Reaction clustering

#### 3.2.3.1 Quality of reaction clusters

Having seen that the EFM-based reaction relatedness approach captures the existence of weakly related reactions, in a way which conventional metabolic pathways fail to, we wanted to check whether the EFM derived hierarchical relatedness linked reactions in a way which makes biological sense. Correlation coefficients were used to hierarchically cluster similarly correlated reactions together (see Methods), and indeed this appears to generate meaningful associations between reactions at different levels of relatedness, in ways which are not possible by considering the Aracyc database directly.

Some examples of close associations, (between strongly correlated groups of reactions), derived from EFMs, but not possible when considering reaction relatedness at any linkage distance in the Aracyc database are

$$\text{Urate} \xrightarrow{PWY-5691} \text{Allantoin} \xrightarrow{PWY-5697} \text{S-ureidoglycolate}$$

which links two complete Aracyc pathways PWY-5691 and PWY-5697, to provide a contiguous route for the degradation of urate, with production of urea, and

$$\text{5-phospho-}\alpha\text{-D-ribose-1-diphosphate} \xrightarrow{PWY-6277} \text{5-phosphoribosyl-5-aminoimidazole}$$

$$\text{5-phosphoribosyl-5-aminoimidazole} \xrightarrow{PWY-6124} \text{inosine-5'-phosphate}$$

$$\text{inosine-5'-phosphate} \xrightarrow{PWY-7219} \text{Adenosine-mono-phosphate}$$

which links pathways PWY-6277, PWY-6124, and PWY-7219 to provide a coherent path for the synthesis of Adenosine-mono-phosphate from 5-phospho-$\alpha$-D-ribose-1-diphosphate. These examples also illustrate that in general, the derived clusters tend to gather successive reactions to form discrete subnetworks of the metabolic system, allowing fairly straightforward biological interpretation.

When the distance threshold for clustering is relaxed, allowing more distantly related clusters to be grouped, EFM derived clusters still make biological sense, but reflect more distant biological relationships, for example linking reactions associated with the photosynthesis light reactions, Calvin-Benson cycle, gluconeogenesis, oxaloacetate shuttle, and sucrose and chlorophyll synthesis pathways into a single cluster.

Importantly, Aracyc pathways are also decomposed, for example, at no linkage threshold distance (prior to formation of the root cluster), are all reactions of the Aracyc 'Calvin-Benson cycle' grouped together using the EFM correlation approach. This reflects the important position of these reactions in the centre of carbon metabolism and their role in the interconversion of many different metabolites, not just in photosynthesis. This indicates that the subjective

judgements historically used to assign to particular groups of reactions to pathways may reflect one of their 'functions', but do not always reflect a reaction's versatility, and therefore the co-occurence of a group of reactions across steady state flux modes.

To avoid cherry picking example clusters derived from this analysis of EFMs, and to assess the general quality of the reaction clusters returned, we evaluated the enrichment of gene ontology terms within each EFM derived cluster. Figure 3.8 shows that clusters generated using correlation coefficients through EFMs, group reactions with similar gene ontology annotations together, across the three independent gene ontologies, and that this occurs to a statistically unlikely extent. This indicates that in general, this method is sensibly grouping biologically related reactions together, across all threshold distances considered.

The reaction relationship data exposed through the EFM derived clusters is of course contained somewhere within the Aracyc data, as after all, the Arabidopsis model used is largely derived from it. However it is not explicitly available. We have previously seen that EFM derived reaction relationships capture information not in pathway databases, and now demonstrated that the same information can be used to group the reactions of a genome scale model of Arabidopsis in a manner consistent with biological intuition, purely via analysis of the structure of the network in an automated process, and that methods using EFMs are one way of doing this.

#### 3.2.3.2 Comparison to transcriptomic derived clusters

It is interesting to compare predicted pathways to those derived from biological data to see whether the historical paths of the Aracyc database, or the EFM derived groups more accurately reflect to reactions which are actually co-expressed in the system.

Here we use correlated transcript expression across the 1,000s of microarray experiments stored in the Expression Atlas database to cluster genes, in order to see whether the gene products associated with similar groups of reactions co-occur in this dataset to those associated in the Aracyc database, or derived through the EFM analysis.

Figure 3.9 shows that neither method seems to group reactions similarly to the groups based on gene expression correlation. This is shown by the relative similarity between the quality distributions of the 'real' reaction groupings, and the randomised ones, which is seen at all linkage distances. In fact the traditional Aracyc pathways seem to be slightly more related to the gene expression data than the unbiased EFMs pathways. This is perhaps unsurprising given than the manual assignation of pathways is ultimately generally based on experimental evidence that they have some co-function. Therefore although, (as we have previously seen), these manually identified pathways can be expected

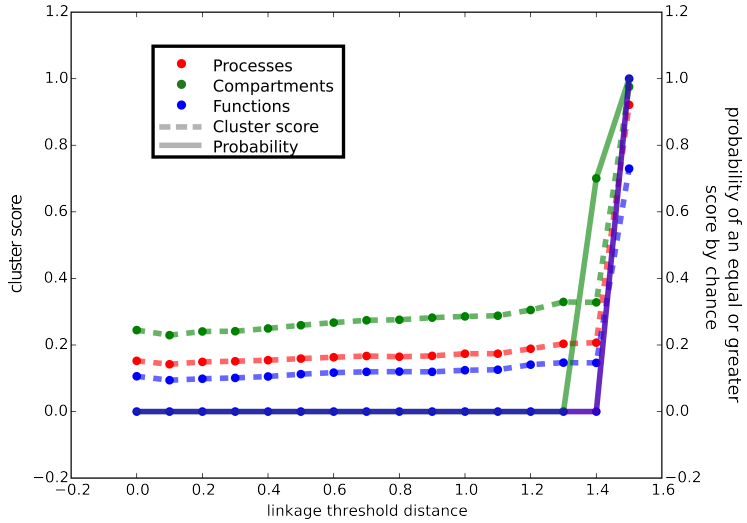



Figure 3.8: Reactions which are clustered together are enriched for the same gene ontology annotations, suggesting that the clusters are generally biologically meaningful. Dashed lines indicate cluster score for enrichment of the same gene ontology terms within each cluster, solid lines indicate the probability of achieving an equal or greater cluster score by chance given the underlying cluster distribution. We see that the clusters generated through EFM analysis tend to group reactions with the same gene ontology terms together to a statistically unlikely degree until the linkage threshold reaches 1.3 or 1.4. At this distance, the number of clusters becomes too small to allow statistically significant enrichment. Gene ontology terms were not used in the clustering method, and so this indicates that the clusters derived through EFM analysis are recapturing biological information, and grouping reactions sensibly. See Methods for the calculation of the gene ontology annotation enrichment within clusters score.

not to tell the full story, by identifying all of the 'functions' of a reaction across metabolic conditions, and therefore miss information about weakly related sets of reactions, they should correspond to groups of reactions which are strongly related in at least one function, and therefore commonly co-expressed. I.e. the manually curated clusters can be expected to have a high accuracy, relative to the transcriptomic clusters, even if they may have a low recall for relatively weakly related 'superpathways'.

Flux control is not homogeneously distributed throughout all reactions of a metabolic network, but is often higher at branch points [120]. We considered that enzymes which catalyse these reactions might be more likely to be tightly regulated, and that therefore gene expression might be more strongly correlated with reaction flux. We therefore segregated reactions by position within pathways according to the Aracyc database in order to see whether this improved the similarity of the reaction groups generated from EFM and transcriptomic data. This segregation used the positional information in the Aracyc pathway, as cluster generation using the EFMs method does not guarantee that contiguous reactions are grouped together (such that the product of one reaction is the substrate for the next), and so it is unclear which reaction is upstream of which. Figure 3.10 shows only a small difference in the similarity of clusters generated from these subsets with the transcriptomic data. Surprisingly, 'middle' reactions (not at the start, or branch points of pathways) appear to be grouped slightly more consistency with the experimental data.

We also considered the similarity of clusters featuring reactions with particular gene ontology annotations (see Methods). A surprising number of relatively high performing annotations are associated with the plastid and mitochondria (Table 3.1). One cause of disagreement between calculated, and transcriptomic reaction clusters is likely to be the strong disconnect between transcript abundance and flux through a reaction [194]. Correlation between transcript and protein abundance is well known to be weak, and flux through associated reactions is likely further disconnected due to post translational regulation, and dependence upon the concentration of reaction substrates and products - that is, indirect regulation by the regulation of up, and downstream reactions.

In contrast to eukaryotes, where post-transcriptional regulation of genes is widespread, post-transcriptional regulation is relatively uncommon in prokaryotes [96]. It is tempting to speculate that the evolutionary origin of plastid and mitochondria allows greater correlation between transcript abundance, and reaction flux, and this is the cause for this enrichment.

### 3.2.4 Conclusion

The degree to which metabolic networks can be usefully modularised is unclear [170, 175], but this has not limited the pragmatic, and well established use of pathway concepts to impose order on the interpretation of reaction net-

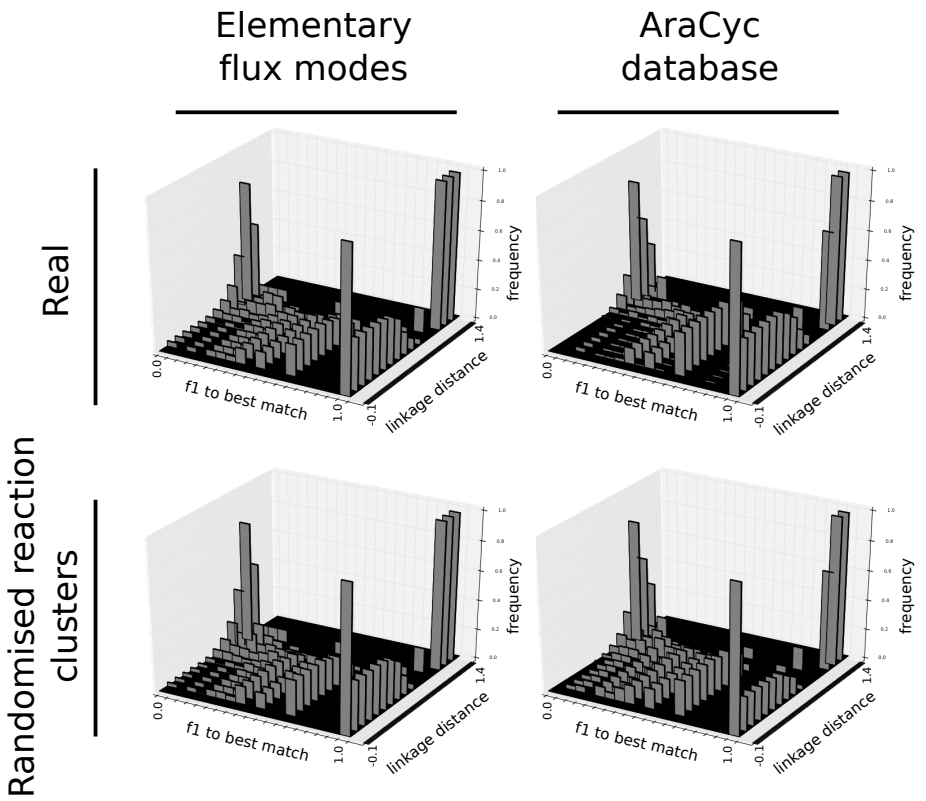



Figure 3.9: Neither the Aracyc database, nor EFMs analysis leads to similar reaction clusters as the transcriptomic data. Cluster similarity was calculated and comparison made as described in Methods. Aracyc and EFMs derived clusters, were compared to those derived from the transcriptomic experiments in the Expression Atlas database. For each linkage distance considered for the Aracyc, of EFMs data, comparison is only shown to the transcriptomic clusters generated using a linkage distance which gives the most similar total number of reaction clusters. The distribution of $F_1$ distances calculated for each linkage distance are shown. In the bottom row, reaction identities were randomly assigned to clusters following the same cluster size distribution as the real clusters generated using either method, in order to indicate the level of similarity forced by the underlying cluster structure. This randomisation shows that, for example, at the largest distances, 1.4, $F_1 = 1$ for all clusters, as at this linkage distance there is only a single cluster comprising all reactions.This shows that for the Elementary modes, the $F_1$ distributions are similar in the real, and randomly assigned clusters, indicating that this method does not generally group reactions together in a more accurate manner than randomly assigning them, given the underlying cluster structure. This does not mean that small subsets of reactions could be accurately grouped, but they are obscured by the generally poor grouping. Conversely, for the Aracyc database, we see a small improvement between the random, and real groups.

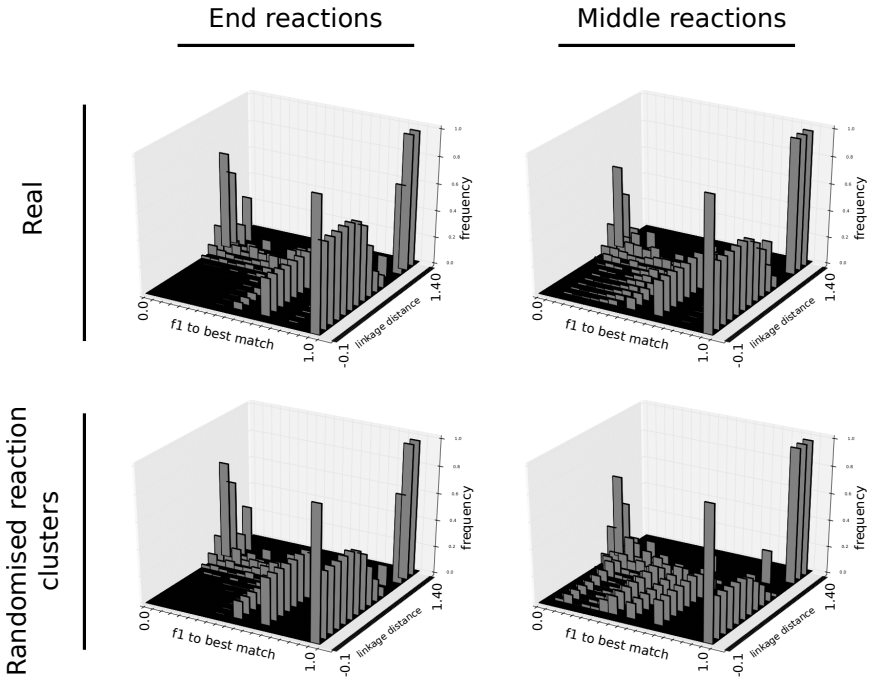



Figure 3.10: Segregating reactions by their position in metabolic pathways did not lead to clear improvement in similarity to transcriptomic clusters. Reactions were partitioned into 'end', and 'middle' reactions according to Aracyc database defined paths. End reactions are reactions which are at the start, end, or branch points within paths, middle reactions are all other reactions. Only reactions which can be mapped between the genome scale model, the Expression Atlas database, and the Aracyc database were included. Comparison was carried out as described in Methods, and Figure 3.9. We see that neither the 'end' nor 'middle' reactions perform much better than random, given the underlying cluster size distributions.

Table 3.1: Reactions associated with organelle gene ontologies are grouped more similarly by EFMs, and transcriptomic data than other reactions. For each gene ontology term, all reactions with that annotation were identified. For each identified reaction, the $F_1$ similarity between the EFM derived cluster containing it, and the most similar transcriptomic derived cluster was calculated. Each gene ontology annotation was scored as the mean of these similarities. The probability of randomly finding better a performing reaction sets was numerically estimated by drawing 100,000 random sets of reactions of the same size as the number associated with the gene ontology, and finding the fraction of these sets with a greater mean $F_1$ score than the reactions associated wth each gene ontology. Corrected probability is the estimated probability, corrected for multiple gene ontology comparisons by the Bonferroni method, in which the estimated probability is multiplied by the number of comparisons made (see section 3.4). Here we show the annotations which perform better than expected (estimated probability cutoff $\leq 0.05$). We see that gene ontologies associated with the chloroplast, and the mitochondria are over represented, meaning that EFMs group clusters featuring these gene ontologies more similarly to transcriptomic clusters than on average. This might be because of greater correlation between transcript abundance, and reaction flux for these gene products than other genes.

| Gene Ontology annotation | Estimated probability | Corrected probability |
|---|---|---|
| *Chloroplast* | | |
| chloroplast thylakoid lumen | 0 | 0 |
| chloroplast thylakoid membrane | 5.59E-03 | 6.76E-01 |
| chloroplast stromal thylakoid | 0 | 0 |
| plastid thylakoid membrane | 1.00E-05 | 1.21E-03 |
| integral component of chloroplast outer membrane | 4.96E-02 | 6.01E+00 |
| chloroplast inner membrane | 1.67E-03 | 2.02E-01 |
| chloroplast | 0 | 0 |
| plastid chromosome | 3.90E-04 | 4.72E-02 |
| plastid | 1.90E-03 | 2.30E-01 |
| plastoglobule | 1.20E-03 | 1.45E-01 |
| photosystem I | 3.80E-04 | 4.60E-02 |
| photosystem II oxygen evolving complex | 0 | 0 |
| PSII associated light-harvesting complex II | 0 | 0 |
| photosystem II reaction center | 1.00E-05 | 1.21E-03 |
| chloroplast ATP synthase complex | 0 | 0 |
| magnesium chelatase complex | 1.83E-02 | 2.22E+00 |
| *Mitochondria* | | |
| mitochondrial inner membrane | 2.00E-05 | 2.42E-03 |
| mitochondrial respiratory chain complex II | 1.02E-02 | 1.23E+00 |
| carbamoyl-phosphate synthase complex | 1.00E-05 | 0.00121 |
| *Other* | | |
| nitrite reductase complex | 0 | 0 |
| glycerol-3-phosphate dehydrogenase complex | | |
| extracellular vesicular exosome | 0.012 | 1.452 |
| vacuolar membrane | 0.00626 | 0.75746 |

works. We have seen that the analysis of elementary modes which include the biomass reaction produce significantly different reaction correlation results to previous methods which the full set of EFMs. We have also shown that the distribution of reaction correlation coefficients thus generated is more similar to experimental data than those from traditional pathway databases. Furthermore the hierarchical clusters generated from these coefficients make intuitive biological sense.

However, we have seen that related reaction clusters derived from both the Aracyc database, and EFMs bear little similarity to clusters derived from transcript expression profiles. This is likely to be partially caused by the use of transcriptomic data as a proxy for true flux measurements, rather than true flux data, but also due to some assumptions made in the comparison.

In deriving EFM clusters from the correlation of reactions through all calculated modes, we have assumed that each of these flux modes is equally likely to be represented in the flux patterns expressed under experimental conditions in the Expression Atlas database, and therefore all EFMs were weighted equally when calculating reaction correlation coefficients. However, this is unlikely to be true. Only a relatively small subset of EFMs are likely to contribute to a given steady state flux [12]. In our naive approach, we had hoped that by considering thousands of such experimental states, the plant may be forced to utilise a wide array of biomass producing modes and therefore we might see similar behaviour between transcript behaviour and the full EFM set, however this has not been the case.

A number of methods have been developed to attempt to identify the principle elementary modes contributing to a given experimental flux state [239, 153, 236]. It is possible that a strategy which focuses on more biologically relevant modes, either through incorporating, for example, more sophisticated thermodynamic constraints [64], or in which correlation of reaction flux is weighted towards these principle elementary modes might yield more similar clusters to the transcript data.

We have not explicitly considered the experimental conditions used in the transcriptomic experiments. There is some evidence that related reaction modules alter depending on environment [247], and so we may be too ambitious in trying to find consistent modules across a broad array of experimental conditions. It is possible that a better approach to finding interpretable reaction groups is to accept metabolic flexibility, and to define highly related reaction groups as pathways, but to allow reactions to belong to multiple groups, and to make explicit the scenarios in which membership of a particular group can be expected to co-occur. In this case, similar gene expression experiments should also be grouped, prior to any comparison with EFMs.

We have seen the potential of EFMs for the study of reaction function and co-expression. However, the final approach still requires some methodological optimisation, either in terms of the method itself, or in the processing of data

its results are compared to.

## 3.3 Nutrient requirements

### 3.3.1 Introduction

We have seen that a reaction modularisation approach cannot easily be used to facilitate the analysis of the calculated EFMs. However, in addition to modularisation, other simplification methods exist for the analysis of EFMs. Here we use a yield space [207, 89] based approach to analyse the nutrient requirements for biomass production. In our approach, a slice through the hyper-cone of possible metabolic states is considered which produces unit biomass. This slice is projected onto a two dimensional (reaction) axes in order to facilitate its simple visual analysis. In contrast to previously published work using this strategy, rather than considering the effect of altered reaction flux on yield, we examine the effect of nutrient uptake on the requirement for other nutrients. This allows us to analyse nutrient requirement phenotypes, and potential nutrient requirement trade-offs described by the calculated EFMs, and the reactions which exhibit significant control over these relationships.

Flux balance analysis has previously been used to consider some aspects of nutrient requirement tradeoffs [168], however, as discussed in chapter 1, these studies only consider some 'optimal' subset of fluxes, and results are highly dependent on the objective function used, which is assumed to be the objective that the network is regulated towards achieving. It is not clear that commonly used objective functions, such as the minimisation of total flux, or the maximisation of biomass yield are entirely appropriate under any biological circumstances, as it is likely that networks are regulated so as to manage a trade-off between objectives [187]. Furthermore, it seems likely that any objective will be dependent upon tissue type, and environment, for example under starvation conditions it seems more likely that efficient networks will be favoured, whereas under favourable conditions flux distributions which allow rapid growth may be preferred.

Therefore we analyse the capabilities of the nutrient requirement spaces determined only by the structure of the reaction network itself, as the results of this approach are likely to be more robust across environmental and tissue differences. As discussed in chapter 1, tissue specific models are made by considering subsets of reactions from the full model, and therefore any tissue specific flux solutions must be subsets of this full nutrient requirement space. That said, our approach essentially considers the whole plant, and does not allow for universal conclusions irrespective of tissue type and developmental stage. We only consider elementary modes which are able to produce metabolites in the ratio required for biomass on average. Consequently, for example, our approach is unlikely to be informative about metabolism in the mature leaf, which is no longer

gaining biomass, and instead primarily producing intermediate metabolites for export to the rest of the plant.

We do not attempt to estimate internal reaction fluxes. Although we expect that the 'correct' flux solution (when averaged over the whole plant) is contained somewhere within the predicted metabolic space, determining exactly where requires accurate measurements, either of nutrient uptake [244], or transcriptomic data [145] to further constrain the space. The structure of biological reaction networks tends to permit degenerate solutions, which cannot be differentiated with even large measurement sets, and require the assumption of some objective function to resolve completely. Given that it appears that even simple organisms strive to balance competing objectives [187], this is not ideal.

Here, we discuss the use of two metrics for evaluating nutrient use efficiency in elementary modes. *Uptake flux* is based on flux though exchange reactions with the environment, (nutrient uptake reactions), in a manner similar to that used for the analysis of energy use efficiency (as, for example [6]). The other, *element flux* is based on total flux of each element through the whole elementary mode, in a manner more similar to that previously used for estimating the total enzyme requirements in constraint-based models, based on total flux through all reactions (as, for example [33]).

We analyse the relationships between various nutrient requirements. Although a number of previous studies have been concerned with the efficiencies of plant metabolic networks, they have primarily studied energy, and carbon conversion efficiency. Here we consider the relationships between nutrient requirements for a larger number of minerals essential to plant life. This potentially has practical applications for better understanding for example nutrient fertiliser requirements, as well as the fundamental constraints on nutrient efficiency strategies and trade-off options available to the plant imposed by the structure of the reaction network itself. We then attempt to identify important reactions which can be used to predict, and potentially modify plant nutrient requirements. We also examine the consequences of environment on metabolic flexibility, and provide a flexibility based hypothesis to explain the form in which nitrogen is predominantly taken up by plants.

### 3.3.2 Analysis using Uptake flux

It is intuitive to consider the nutrient requirements of a flux distribution as directly related to flux through the nutrient uptake reactions from the environment into the organism. These reactions in the considered model are shown in Table 3.3. Flux balance analysis studies have commonly applied various metrics based on this intuition to energy use, in which flux through some reaction importing 'photons', or high energy metabolites into the organism, is divided by an output flux producing some product of interest in order to calculate an efficiency measure [6, 183]. This approach has also, although less frequently, been

taken to evaluate carbon, and nitrogen use efficiency in plants [82, 35, 5].

In direct analogy to these metrics, here we analyse the nutrient requirements of an elementary flux mode by considering the amount of flux through each nutrient uptake reaction required to allow unit flux through the 'biomass' reaction.

Table 3.3: Nutrient uptake reaction equations. Metabolites to the right of the arrow are imported into the modelled organism, metabolites to the left are exported.

| Identity | Formula |
|---|---|
| CO2 tx_f | $\Longrightarrow$ Carbon-dioxide [c] |
| CO2 tx_b | Carbon-dioxide [c] $\Longrightarrow$ |
| O2 tx_f | $\Longrightarrow$ Oxygen [c] |
| O2 tx_b | Oxygen [c] $\Longrightarrow$ |
| Photon tx | $\Longrightarrow$ Photon [c] |
| GLC tx | $\Longrightarrow$ Glucose [c] |
| NH4 tx | $\Longrightarrow$ Ammonia [c] |
| NO3 tx | $\Longrightarrow$ Nitrate [c] |
| Pi tx | $\Longrightarrow$ Phosphate [c] |
| SO4 tx | $\Longrightarrow$ Sulfate [c] |
| Ca tx | $\Longrightarrow$ $Ca^{2+}$ [c] |
| Fe tx | $\Longrightarrow$ $Fe^{2+}$ [c] |
| K tx | $\Longrightarrow$ $K^{+}$ [c] |
| Mg tx | $\Longrightarrow$ $Mg^{2+}$ [c] |

Figure 3.11 shows correlation between flux through these nutrient uptake reactions across elementary modes. We see strong internal correlation between those nutrient uptake reactions associated with an autotrophic lifestyle, (Photon tx, CO2 tx_f, O2 tx_b), those with a heterotrophic lifestyle, (GLC tx, CO2 tx_b, O2 tx_f), and strong anti-correlation between these two groups. This supports both the proposed model of Arabidopsis, and also this approach to understanding nutrient requirements, as it reflects basic experimentally determined relationships between these fluxes.

We additionally see a number of interesting results which apparently merit further investigation, suggesting that this approach generates non-obvious questions. For example what causes the strong positive correlation between potassium and calcium uptake, and their mutual negative correlation with regards to magnesium? Or the weak association between increased nutrient uptake and autotrophy? Or most strikingly, the strong positive correlation between ammonium, phosphorus, and sulfur uptake, a group which intriguingly does not include the other nitrogen containing nutrient, nitrate?

In investigating this final conundrum, we produced Figure 3.12, in which several nutrient efficiencies of the calculated elementary modes are plotted against each other. Each set of axes corresponds to a different two-reaction projection of the EFMs. The same EFMs in different axes are linked by grey lines. The nutrient requirement space which is predicted to be available to the plant consists of

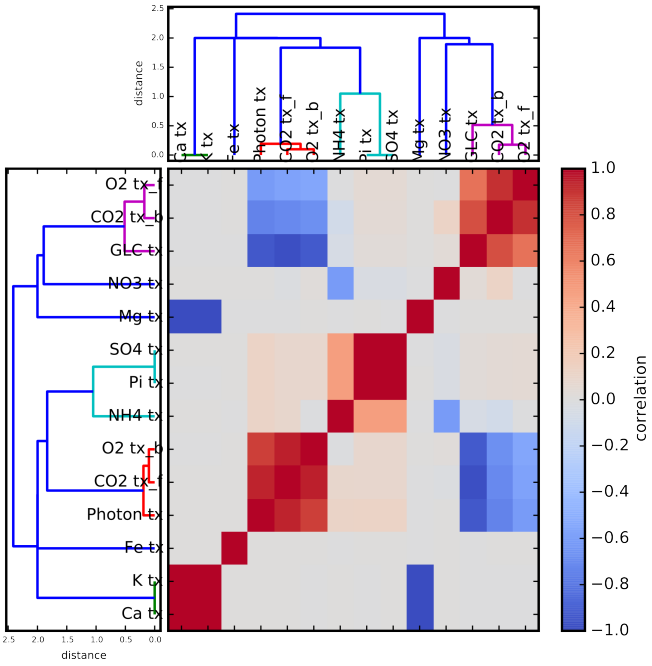



Figure 3.11: Correlation coefficients between nutrient uptake reactions. Clustering by average linkage distance. Nutrient exchange reactions associated with autotrophy and heterotrophy are grouped together due to mutual correlation within each group, and anti-correlation between the groups. This is as expected, and validates the model.

linear combinations of elementary modes, and therefore the permissible nutrient efficiency space is the convex hull of the plotted points.

Although 300,000 elementary modes are plotted, they form extremely discrete clusters in Figure 3.12, and in fact are largely superimposed upon each other. Additionally, although four nutrient uptake phenoypes are plotted (flux through $SO_4$ tx, Pi tx, $N_{tot}$ tx, and the fraction of $N_{tot}$ taken up un the form of ammonium), we can see that most of the differences across these phenotypes can be largely explained in terms of elementary mode membership in only four groups. Group 0 vs group 1 explains Pi tx variation, almost all $SO_4$ tx variation, and all variation in $N_{tot}$ for EFMs which uptake some nitrate. Group 2 vs group 3 captures an extremely small difference in the amount of sulfate required by different elementary modes.

Yield space analysis is known to drastically reduce the number of elementary modes which must be considered in order to explain the behaviour of the system [207], however it is surprising that there should be so few factors underlying the different nutrient requirements of so many elementary modes. Figure 3.12 also shows that nutrient tradeoff relationships are apparently surprisingly highly constrained. For example, it suggests that the structure of the reaction net-

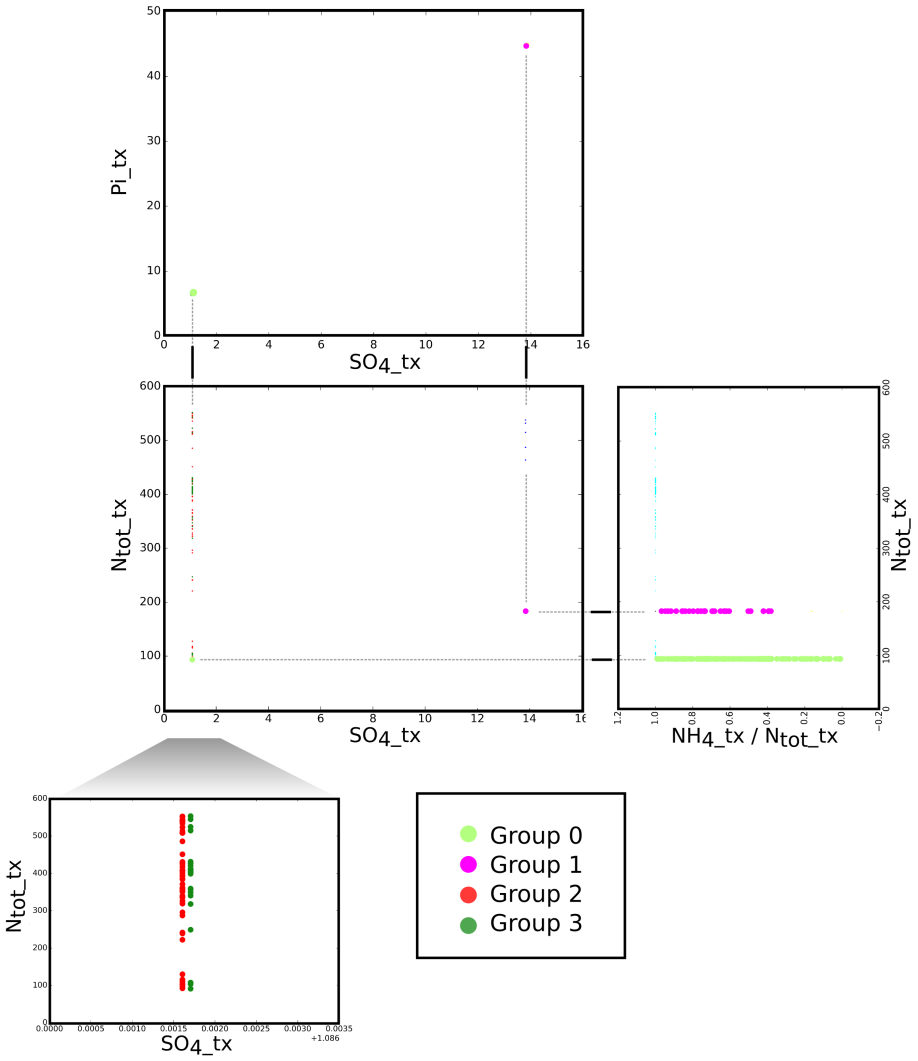



Figure 3.12: Uptake reaction flux phenotypes. Flux through indicated nutrient uptake reactions, normalised by flux through the biomass reaction for the first 300,000 EFMs found in the Arabidopsis model. $N_{tot}$ tx is the sum of flux through the $NH_4$ tx and $NO_3$ tx reactions. Dashed grey lines highlight the same elementary modes projected onto different axes. Predicted nutrient requirement phenotype are surprisingly discrete, and apparently linked, indicating that a small number of differences in the underlying reaction flux distribution cause the different phenotypes seen.

Table 3.4: Empirical formulas of metabolites in the reaction *RXN-5984*. This unbalanced equation is the cause of the Group 2 versus Group 3 sulfate uptake phenotype.

| metabolite | C | H | N | O | S |
|---|---|---|---|---|---|
| dethiobiotin | 10 | 17 | 2 | 3 | - |
| S-adenosymethionine | 15 | 23 | 6 | 5 | 1 |
| **substrates total** | **25** | **40** | **8** | **8** | **1** |
| | | | | | |
| S-adenosyl-L-homocysteine | 14 | 20 | 6 | 5 | 1 |
| 9-mercaptodethiobiotin | 10 | 17 | 2 | 3 | 1 |
| **products total** | **24** | **37** | **8** | **8** | **2** |

work means that the requirement for sulfate, and inorganic phosphate must be (almost) perfectly correlated, with no metabolic flexibility available.

To understand what difference in the flux distributions 'causes' the separation of EFMs in groups 0 & 1, and 2 & 3, we manually examined all reactions which carry the same flux in all EFMs in group 0 but not 1, and *vice versa*, and group 2 but not 3 and *vice versa*.

The separation of groups 2 and 3 is caused by a different route for the production of Biotin. Elementary modes in Group 2 all carry flux through RXN-5984 which contains a mass balance error (see Table 3.4), leading to greater sulfur use efficiency. This reaction was present in the original, published model [33], from which the current model was developed, and likely originated due to an erroneous Aracyc entry, which has since been corrected in the current database version.

The elementary modes in group 1 produce 'excess' coenzyme-A, which is then exported from the model, resulting in greater flux through the export reaction in group 1 relative to group 0. The empirical formula of coenzyme-A is $C_{21}H_{36}N_7O_{16}P_3S$, which links the greater group 1 requirement for sulfate, inorganic phosphate, and nitrogen qualitatively, although it does not entirely capture the quantitative differences in the requirements of these nutrients (Figure 3.12). It is therefore possible that some secondary metabolite(s) are also exported, and differentiate these groups. However, export of coenzyme-A can be seen to be necessary and sufficient for export of any other associated metabolites, and therefore all differences between group 0 and group 1 can be considered as a single phenotype.

Determining cause and effect of reaction fluxes of an elementary mode can be somewhat circular, but we wanted to understand the underlying metabolic difference between elementary modes in group 0 and 1 which necessitates the export of coenzyme-A. However, whilst elementary modes within groups 0 and 1 are identical within the projection shown in (Figure 3.12), each elementary mode is unique when all reactions are considered. Unfortunately, the presence of complex networks of mutually compensatory, essentially parallel paths made

it impossible to trace the underlying cause of the difference, beyond the network shown in Figure 3.13. However, it seems likely that some conserved difference in flux through central metabolism causes the difference seen between groups 0 and 1.

The difference between considering energy use efficiency, and nutrient use efficiency is that whereas chemical energy can be lost by reaction transformations, most published metabolic models are mass-balanced, such that that atoms cannot be lost, or produced by reactions. This means that different reaction flux vectors for the production of the same products from the same substrates can require different amounts of energy, but cannot require different amounts of nutrients to be taken up into the model. Instead, we have seen that within this analysis framework, different nutrient use efficiencies (uptake fluxes) arise via errors in reaction stoichiometry, and (predominantly) via the differing production of 'waste' metabolites, as a consequence of the reaction route taken to produce the biomass reaction substrates. 'Waste' metabolites are considered to be those which cannot be recycled into the production of more biomass substrates and are exported from the model. In the biological system they are excluded from metabolism, either by excretion from the plant, or transportation into the vacuole for indefinite storage.

This suggests a large sensitivity of the nutrient phenotypes seen, to the particular metabolites which are permitted to be exported from the model. For example, it is possible that multiple underlying reaction flux differences cause the nutrient requirement differences between group 0 and 1 elementary modes. It is possible that the constraints of the model subsume them into a single 'phenotype' through the need to consume all nutrients in the correct ratio required for coenzyme-A production, as the 'true' waste metabolites cannot be exported, but are instead converted to coenyzme-A.

This hypothesis predicts different sensitivities of various elements to the permitted waste metabolites. Relatively common elements, such as C, are present in almost all permitted waste metabolites, and so, disposing of excess C can be achieved in many ways, resulting a relative insensitivity to the permitted waste metabolites. Therefore phenotypes involving carbon requirements, (for example, the correlation between heterotrophic and autotrophic uptake fluxes shown in Figure 3.11), are robust to the permitted export metabolites. Conversely, relatively rare elements, such as sulfur, phosphorus and nitrogen are in fewer exportable metabolites. In this case, any intermediate waste metabolite containing phosphorus must be converted to coenzyme-A, with the concurrent recruitment of sulfur, leading to their highly correlated relationship. We also see extremely discontinuous rare nutrient requirements in Figure 3.12. This is possibly a consequence of the complex, interacting stoichiometric requirements of reaction networks lead to 'rounding-up' the nutrient use inefficiencies into discretised levels.

The addition of alternative waste metabolites is expected to break the mutual dependency of rare elements for the production of coenzyme-A, and therefore

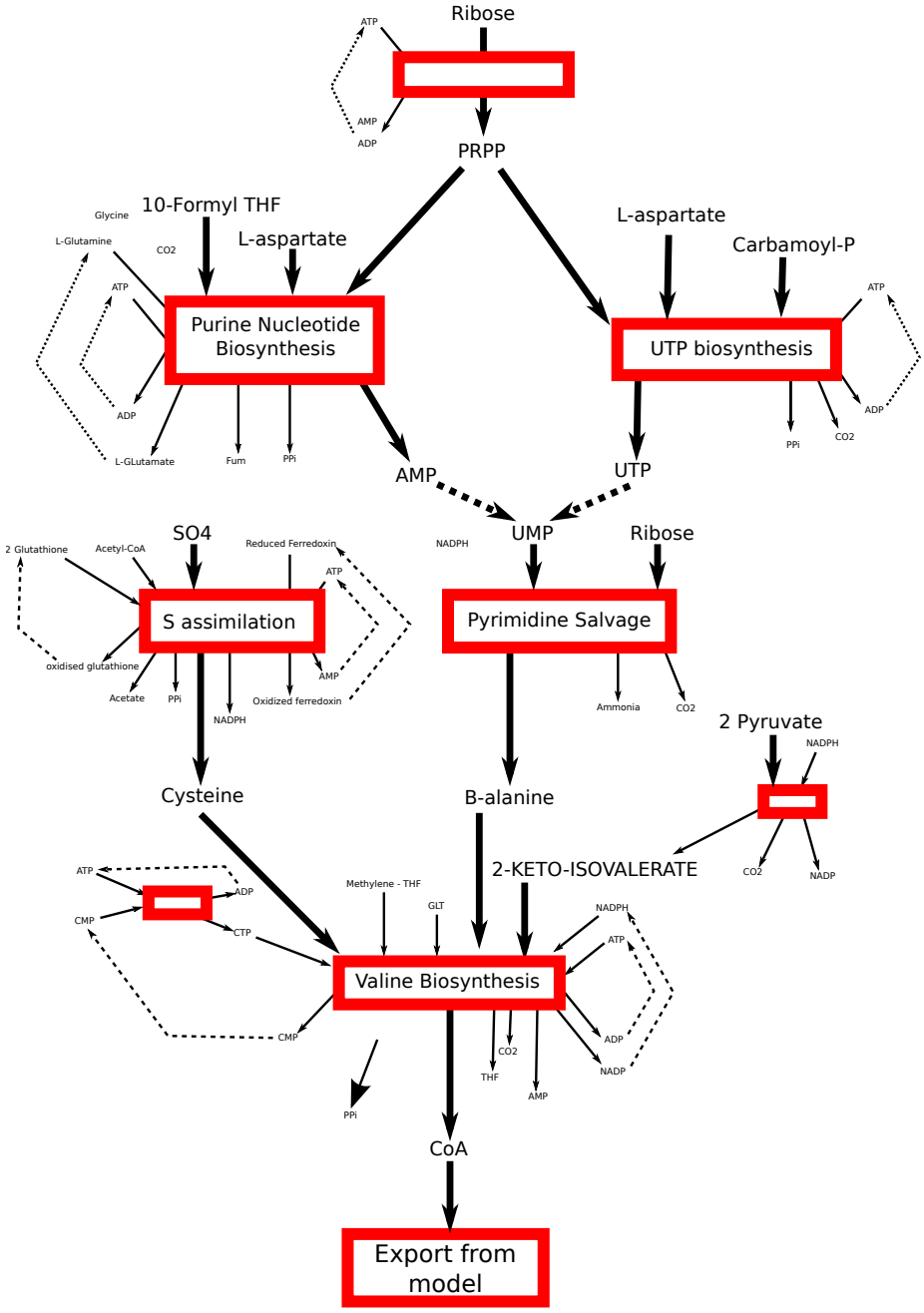



Figure 3.13: The metabolic processes which differ between Group 0 & Group 1, and which are therefore related to the underlying cause of the nutrient uptake phenotypes. Red boxes indicate amalgamated reaction groups, metabolites going into each box indicate net substrates of the grouped reaction, metabolites coming out are net products. Dashed arrows indicate obvious recycling of metabolites, Large font metabolites are substrates taken up directly from the environment (SO4), or candidate metabolites which underlie the need for the export of CoA (coenzyme-A).

also break the correlation between uptake fluxes. This idea is corroborated by the disappearance of the discussed EFM phenotype groups in a modified model in which the individual export of all components of the 'biomass' equation was permitted, (here called the 'flexi-model'). Most importantly, some of the predicted overall nutrient requirement relationships disappear completely, for example all correlation between $SO_4$ tx and Pi tx, disappears.

The observed sensitivity to exportable waste metabolites can be addressed by allowing only the export of the correct (biological) waste metabolites, or by considering a different nutrient use metric. It is not necessarily easy to identify waste metabolites experimentally, as the plant vacuole can act as both a permanent bin for 'true' waste metabolites, but also as a storage site for metabolites, which are later recycled back into active metabolism. Furthermore, given the quality of many genome scale metabolic reconstructions, it is not clear that all true waste metabolites can be produced in many models. These difficulties are possibly reflected in the relatively little discussion in modelling papers afforded to the waste metabolites as compared to metabolites which are allowed to be taken up, and to the derivation of the biomass equation. Therefore although uptake based efficiency metrics can be useful to analyse the energy efficiencies of elementary modes, and to a lesser degree the requirements for common elements, which allows autotrophic, and heterotrophic lifestyles to be captured using uptake based metrics, the nitrogen, sulfur, and phosphorus relationships seen are likely artefacts of model inaccuracies.

We therefore proceed using a different metric which is more robust to the identities of metabolites which can be exported. Termed elemental flux, this is the total flux of an element within a reaction flux distribution (see subsection 3.4.4 for details of calculation). Intuitively, this is the amount of the element which must be present somewhere within the metabolic network at metabolic steady state per unit biomass, and is similar to the common use of total flux as a proxy for the total amount of enzymes required to catalyse a flux distribution.

Whilst the output of this metric is still somewhat influenced by the specific export metabolites allowed, due to the need for reactions to convert between 'true' and exported waste metabolites, it does not suffer from the very discrete nutrient use efficiencies seen when using an uptake flux metric, and in our hands, appears to be more robust to changes in exportable metabolites between the 'fixed' and 'flexi' models. This is likely because these conversions involve relatively small numbers of reactions carrying little flux compared to the whole metabolic network. Although to reduce duplication, we only show results generated using the 'flexi-model', which does not impose permitted ratios on the export of metabolites from the model, and is therefore more robust to potential inaccuracies in the stoichiometry of the biomass equation, the discussed conclusions were also seen using the original model described in chapter 2.

### 3.3.3 Nutrient requirement tradeoffs

More nutrients can be made available to the plant through the application of external fertilisers, however, it is not necessarily clear that it will be able to utilise them, unless additional resources are also made available. Here we evaluate theses mutual nutrient requirements in the biomass producing EFMs. This analysis also allows us to assess the extent to which nutrient requirements can be modulated, and traded-off against each other through favouring alternative metabolic strategies for the production of the same biomass components. Although previous work has shown that EFMs can be used to study the effect of nutrition on plant metabolism [18], this study was concerned with the effect of nitrogen source on internal fluxes, rather than the effect of element requirements on each other.

Figure 3.14 shows the amount of nitrogen, plotted against other elements, required to produce unit flux through the biomass reaction, in autotrophic, heterotrophic, and mixed elementary modes. The plotted points in the 'raw' graphs correspond to the elemental flux of each element in each elementary mode, normalised by flux through the biomass reaction, and so are a measure of nutrient use efficiencies (NUEs). Metabolic flux space is defined by linear combinations of elementary modes, and so the elemental flux space predicted to be accessible to the plant is described by the convex hull of the plotted points under each lifestyle.

Starvation of a nutrient can have the effect either of shifting the flux distribution towards EFMs which are efficient in the utilisation of that nutrient, potentially without compromising biomass production, or of reducing total flux through the system, with an associated reduction in flux through the biomass equation. Positive correlation between nutrient pairs indicates that modes which are efficient in the utilisation of one nutrient are also efficient in the use of the other. As such, starvation of one of a pair of perfectly correlated nutrients is expected to have the same effect as starvation of both, and therefore has the same effect of growth of the plant. Conversely, elementary modes which require more of one element in a perfectly correlated pair cannot be utilised unless the other is also available.

The 'raw' panels of Figure 3.14 show that there is a strong positive relationship between the predicted nutrient use efficiency of nitrogen, and most other elements in the model. The elements shown to be correlated with N are also correlated with each other. Some of the predicted nutrient dependencies agree with previous experimental findings, for example elevated carbon dioxide levels are not taken advantage of in the form of more carbon uptake, unless additional oxygen was available as well [195], and that phosphorus starvation limits carbon uptake [86].

Iron, magnesium and sulfur to not correlate strongly with the requirements for other elements. However, of these, iron and magnesium are not well integrated

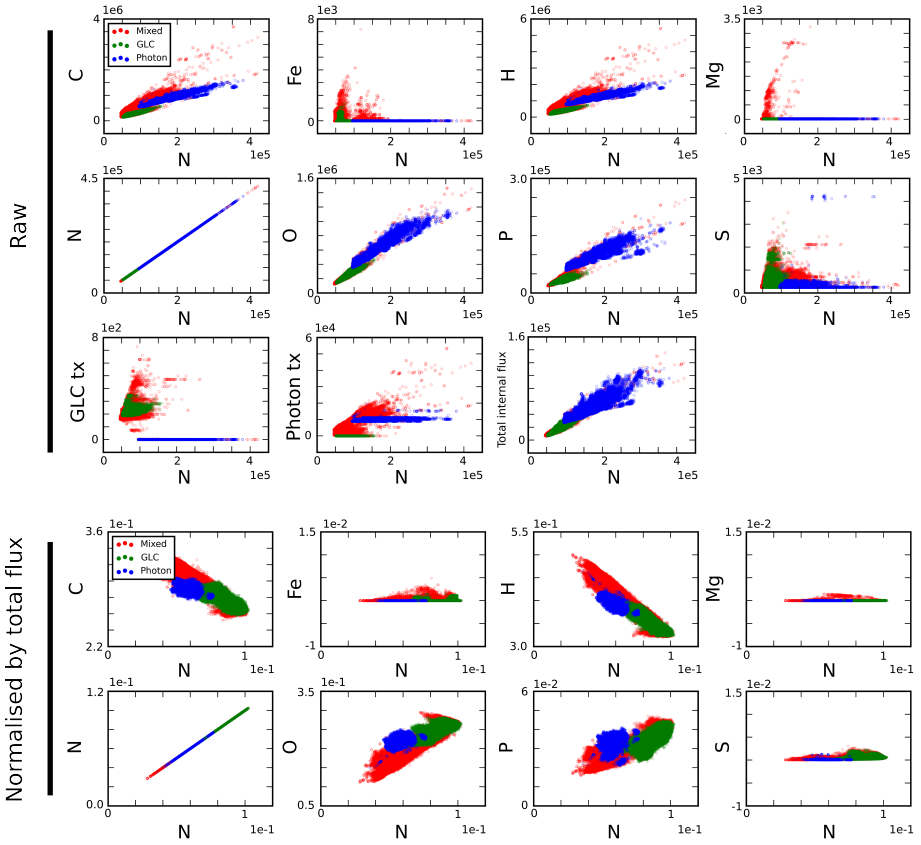



Figure 3.14: Elemental flux relationships. See Methods, for calculation of elemental flux. Each point corresponds to an elementary flux mode, 'Photon' corresponds to autotrophic EFMs in which uptake of glucose is not permitted, 'GLC' corresponds to heterotrophic flux modes in uptake of photons is not permitted, in 'mixed' flux uptake of both energy sources is allowed. In 'normalised by total flux' the 'raw' values are divided by total elemental flux through each elementary flux mode. We see that surprisingly, there appears to be little tradeoff between nutrient requirements in the different metabolic modes. Requirements for many nutrients are strongly correlated, (with the exception of iron, magnesium, and sulfur), indicating that additional access to one resource often does not allow access to alternative metabolic modes, unless other element resources are also available. Correlation between nutrient requirements appears to be largely caused by mutual correlation to total flux, indicating that nutrient inefficient modes are also protein inefficient. 'Mixed' metabolic modes appear to be more flexible than either autotrophic or heterotrophic lifestyles, and are able to access the most nitrogen efficient modes.

into the model; they are present in only a very small number of reactions, and are often associated with proteins, which are not explicitly modelled. We therefore have relatively little confidence in results concerning them.

Although none of the elements are perfectly correlated, we would expect, for example, the impact of mutual starvation of nitrogen and phosphorus to be less severe than the impact of the starvation of nitrogen and sulfur, relative to the starvation of each of these nutrients singly. This is because elementary modes which efficiently utilise nitrogen are also efficient in phosphorus metabolism, and so the extra nutrient stress has little effect in further reducing available metabolic space. In contrast, nitrogen and sulfur stresses are expected to be fairly independent, and so sulfur starvation is expected to further reduce the metabolic space available to the plant.

If we assume that nutrient inefficient elementary modes have some benefit, which is not well captured under the steady state modelling framework, (for example in increased rate of growth), we would also predict that fertilisation with additional nitrogen would not be beneficial unless phosphorus is also sufficiently available, whereas Figure 3.14 shows that most N inefficient modes are comparatively sulfur efficient.

It is not always necessary to assume some hidden benefit for inefficient modes. While we have considered the effects of a shortfall of available nutrients, we have not accounted for any negative or toxic effects from their overabundance. This is likely to be a problem when the primary nutrient stress has reduced the capacity for the integration of a secondary nutrient. For example, although the plant may only be able to incorporate a reduced amount of energy due to other nutrient stresses, incident light will still provide the same amount of energy to the plant. It can be seen in 'N vs Photon tx' in Figure 3.14 that N starvation reduces access to inefficient photon flux modes. Consequently nitrogen starved plants can be expected to be susceptible to photo-oxidative damage. It is, however, not clear that this idea can be extended to toxicity in other nutrients given that mineral concentration in plants is typically greater than in the surrounding environment.

We see correlation between element use and total internal flux, consistent with the experimental finding that increased oxygen levels leads to increased flux throughout metabolism [243]. Although the relationship is clearly more complicated, depending on particular enzyme kinetics, and stability, total flux through a system is often used in constraint-based modelling studies to approximate total enzyme requirements [18]. Figure 3.14 therefore suggests that element efficient modes are also protein efficient.

There appears to be surprisingly little nutrient use efficiency tradeoff between nutrients, energy efficiency, and protein use efficiency in different elementary modes. This raises the question of whether this is an inevitable consequence of being constrained to produce the same molecules in the same ratios to produce biomass, or is a consequence of the particular structure that the metabolic

network has evolved. Reaction networks can easily be conceived which disconnect elemental flux from the elemental composition of the product (for example by introducing superfluous intermediate steps), but only by making them less efficient. This suggests that this inflexibility may be an indication of efficient reaction networks for converting substrates to products. The tradeoffs between efficiency, robustness, and flexibility have been discussed previously [109], however, given the famous robustness of biological networks, it is surprising that the efficient use of elements is so prioritised over their flexible use. Consequently, any presumed tradeoff in the performance of different flux modes is assumed to be primarily at the level of efficient versus rapid growth, rather than in the efficient, flexible use of different resources.

By elimination therefore, alteration of nutrient use efficiency of different elements relative to each other is likely to be achieved primarily by altering the composition of the plant, (that is the biomass reaction equation), rather than by picking between the tradeoffs of different flux modes. It is well known that plants change their morphology and chemical composition under nutrient stresses [203, 5]. It would be interesting to analyse the effect of these alterations changes in yield space. Previous work has found little difference between optimal flux through central carbon metabolism in response to altered biomass composition [244, 168, 249], and concluded that this is likely because production of biomass precursors requires relatively little energy in comparison to 'maintenance flux'. Consequently, while nutrient requirements must change somewhat in response to altered biomass composition (due to the mass-balance constraint), it is not necessarily clear how much they will change. Intuitively more of a nutrient will be required if more metabolites containing it are required for biomass, but it is not clear how great the elasticity with respect to demand for different biomass components will be, or that it will be the same for all elements, when different anabolic routes are considered.

We consider the 'raw' graphs shown in Figure 3.14 to be those most relevant to predictions of nutrient requirements, and the behaviour of the biological system, however, given the strong correlation between all nutrients and total flux, it is interesting to look at the relationships between nutrients in elementary modes when normalised by total flux, which reveals effects obscured by the total flux.

Many relationships do indeed change, this shows that differences in nutrient use efficiency are dominated by differences in total flux, and that this big total flux dependency is what leads to the strong correlation among most elements. The normalised graphs show that there is some scope for the modification of nutrient use efficiency through the use of different flux modes, for example the correlation between nitrogen, and carbon reverses when normalised by total flux, indicating that for a given value of total flux, nitrogen efficient modes are inefficient, and *vice versa*. Conversely, nitrogen, phosphorus, and oxygen requirements are still correlated, although not so strongly, indicating that it is very difficult to separate the requirement for these elements, as they are involved

in the same reactions to a greater degree than elements on average.

Although, as discussed, our results are consistent with some experimental studies, our somewhat simplistic approach does not capture the full behaviour of the system. For example some nutrient stresses should decrease carbon conversion efficiency [137], leading to a greater carbon requirement. Furthermore, contrary to what is indicated in 'N vs Total Internal Flux' of Figure 3.14 experimental nitrogen limitation does not lead to a decrease in total protein levels [225].

In their model of barley seeds, Grafahrend-Belau et al. [73] found only five independent, optimal forms of metabolism as environmental sucrose to oxygen ratios were varied. This reflects the amount of information which is lost in the optimisation step of flux balance analysis, when compared to the thousands of independent forms shown in any panel of Figure 3.14. However considering all elementary modes, as we have approximated here, is probably too broad an approach. We have made the simplifying assumption that the elementary modes calculated are representative of the set used by the plant. However, although plant metabolism must act within the borders of the convex hull of the points presented, it may not utilise the full space, and consequentially could exhibit different nutrient use tradeoffs to those described above. How justifiable this parimonious assumption is is not clear. An interesting next step in this analysis is therefore to link the full elementary modes set calculated to a subset of biologically utilised modes. This can be achieved by the analysis of transcriptomic data under different environmental conditions, and the decomposition of this dataset into the different elementary modes used. In addition to making nutrient tradeoff predictions, comparison between the full elementary set, and used set will also be useful in identifying the 'purpose' behind metabolic regulation, and how general a priority nutrient use efficiency is across environments.

### 3.3.4 Hetero, auto, and mixed; the flexible lifestyles of plants

Here we compare and contrast the constraints on metabolism imposed by autotrophy, (in which energy is provided by photosynthesis), heterotrophy, (in which it is provided by glucose), and 'mixed' metabolism, (in which both are available).

Figure 3.14 shows that elementary modes are on average less nutrient efficient under autotrophy (blue) than in heterotrophic (green), or mixed (red) modes, as autotrophic modes occupy the upper, right-most regions of the plotted space in the 'raw' graphs. Interestingly, there are differences in the cause of the increased nutrient requirements under autotrophy between elements. For example, the normalised 'N vs C' graph shows that carbon is proportionately over-required for autotrophy, relative to other elements. Conversely, heterotrophy is proportionally enriched for fluxes which require nitrogen, and the greater

requirement for nitrogen under autotrophy is entirely caused by the increase in total flux.

The greater total flux in autotrophic modes is largely caused by changes in flux through pathways directly associated with converting photons into chemical energy, but metabolism is generally different in hetero- and auto-trophic modes, with 467 reactions carrying over twice as much flux on average in autotrophic rather than in heterotrophic metabolism. These broad changes are consistent with experimental data which suggests that expression of $\sim 35\%$ of the transcriptome exhibits circadian clock regulation [142]. The 'heterotrophic', and 'autotrophic' models used do not exchange metabolites, and so it is difficult to extrapolate the behaviours of these models to a single system over the diel cycle. However in examining Figure 3.14 it is tempting to speculate that plants may grow primarily at night, rather than day in order to reduce the maximum total flux that must be supported at any single point during the day-night cycle. By dividing metabolism into the production of storage metabolites during the day, and production of biomass during the night, metabolic requirements are smoothed over the 24 hour period, reducing the difference in peak, and trough protein requirements, and potentially allowing increased recycling of amino-acids and cofactors between metabolic pathways at different times of day.

Average nitrogen requirements clearly vary between autotrophic, and heterotrophic flux distributions, being greater overall in autotrophic modes, but greater per unit flux in heterotrophic modes. Interestingly, experimental evidence suggests that the day night cycle is critical for nitrogen assimilation, as the production of glutamine is predominantly fed by the remobilisation of stored molecules from previously assimilated carbon [63]. These results contrast with FBA studies [44], which found no difference in nitrogen requirement in day and night, when using a requirement metric based on flux through nitrogen uptake reactions. This previous finding was likely a consequence of using the same biomass equation in day and night, and the weaknesses of uptake based metrics, as discussed previously in subsection 3.3.2.

It is clear from Figure 3.14 that even a simple difference in carbon/energy source dramatically alters the metabolic capabilities of the organism. Different tissues, which also exhibit different biomass compositions, produce various storage molecules, and potentially have different metabolic 'objectives' as well can therefore be expected to exhibit markedly different nutrient use efficiencies, and metabolic lifestyles.

Figure 3.14 shows that of heterotrophy, autotrophy, and mixed metabolism, mixed allows much greater metabolic flexibility, occupying not only the union space of the other two, and the space accessible through linear combinations of autotrophy and heterotrophy, but also more extreme regions. The increased nutrient use flexibility under mixed metabolic metabolism is due to being able to access different elementary modes. This is likely to be caused by the relaxation of the strict stoichiometric relationships between the products of photosynthesis,

which can be modified by additionally catabolising various amounts of glucose, therefore allowing access to additional elementary modes. The modification of stoichiometric ratios of metabolites produced during photosynthesis has previously been seen to be important, and result in increased metabolic flexibility [6]. Figure 3.14 only shows this flexibility in terms of nutrient use tradeoffs, but it is also confirmed by flux variability analysis, which shows that average reaction flux across the entire metabolic network is also more flexible under mixed metabolism (Table 3.5).

Table 3.5: Reaction flexibility under different lifestyles. Summed flux variability is $\sum r_{max} - r_{min}$ for all reactions, where maximum and minimum values are the largest and smallest fluxes per unit biomass in the calculated set of elementary modes. Mixed metabolism apparently permits greater flux flexibility. Greater metabolic flexibility is often associated with robustness to environmental and genetic perturbation [211].

| lifestyle | summed flux variability |
|---|---|
| Autotrophic | 606,513 |
| Heterotrophic | 231,844 |
| Mixed | 832,448 |

Increased metabolic flexibility allows increased control of the ratios between the different nutrient use requirements of the plant, an indicated by the comparatively broad areas of the graphs that the mixed modes are able to access in Figure 3.14, but also to access the most efficient modes overall. For example Figure 3.14 shows that only through a mixed lifestyle is access to the most efficient nitrogen utilising modes possible.

As discussed in chapter 1, various measures based on EFMs have been used to quantify robustness to genetic perturbation, predominantly based on the fraction of EFMs remaining after a reaction is removed from the network, such that more robust networks generally have more EFMs remaining [173, 242, 16, 13]. Consideration of Figure 3.14 suggests a second metric, in which, rather than assessing the number of remaining EFMs after perturbation, the average volume of the efficiency space remaining is considered proportional to the robustness of the system. This approach also incorporates susceptibility to environmental perturbation. It is expected that mixed metabolism will be found to be more robust under this metric than exclusively auto- or heterotrophic modes, as mixed metabolism starts form a more flexible metabolic position.

The approach we have taken to model these different lifestyles is obviously crude. We are notionally considering the requirements of the whole plants but do not link the models to promote the production of storage molecules during autotrophy, or allow the utilisation of a range of element and energy sources during heterotrophy. Furthermore, we do not alter the stoichiometry of the biomass equation, or account for shuttling of metabolites between roots and shoots in autotrophic metabolism. Other models have more carefully considered metabolism in the whole plant [44], or over the diel cycle [31], however by using

elementary modes, we have derived some interesting general results based on the availability of energy sources, which have not been obvious in these flux balance analysis studies.

We did not constrain the ratio of glucose to photon uptake permitted in the 'mixed' metabolic form. Although in the long term, the plant must produce all available starch, and therefore glucose from photons, effectively putting an upper limit on this ratio, in the relative short term, this is not necessarily the case, due the storage of starch. Additionally, any long term ratio can be expected to vary between different tissues. It is therefore pertinent to consider the effect of an unconstrained ratio. However, it would be interesting to explore the effect of a fixed ratio representative of the ratio over the diel cycle, averaged over the whole plant to see whether this also leads to an increase in metabolic flexibility.

*In vivo*, starch accumulates during the day, and is consumed at night. The rate of starch use at night is carefully regulated, so that $\sim 5\%$ remains at dawn [70]. Overconsumption, leading to the exhaustion of starch stores during the night, is associated with a starvation response, and reduced growth. Conversely however, mutations leading to reduced consumption overnight are also associated with reduced growth [70]. It is therefore unclear why not all starch is consumed by dawn *in vivo*, as this presumably would allow increased growth. Rather than simply an insurance against running out of starch, the advantages of metabolic flexibility, efficiency, and (potentially) robustness, accessible through the mixed metabolism of photons and glucose might explain why it is that although plants use most of their starch stores overnight, under most environments, starch is not completely depleted, even under severe carbon-limitation [165], potentially allowing mixed metabolism throughout the day.

### 3.3.5 Nitrate & ammonium uptake ratio limits scope for metabolic inefficiency

Plants take up nitrogen from the environment as nitrate, ammonium, and to a lesser extent, as proteins or amino acids. Using inorganic nitrogen in the form of ammonium rather than nitrate should be more efficient , due to the large reductant requirement for conversion of nitrate to ammonium prior to integration into organic molecules. This makes up most of the reductant requirement of the cell [138]. The expected difference in energy efficiency between the use of these N sources was initially confirmed by flux balance analysis [82, 44]. It is therefore surprising that plants preferentially uptake a mix of nitrate and ammonium [138].

Retardation of growth through excess ammonium uptake is called 'ammonium toxicity', however the cause of this toxicity is not clear. Ammonium toxicity has been linked to the uncoupling of proton gradients across membranes [20], or futile cycling of the import and export of ammonium [24]. If ammonium toxicity

is caused entirely by these processes, then its analysis is beyond the scope of this modelling framework, however alleviation of ammonium toxicity by nitrate [180, 75] suggests that these may not be the only causes of the effect. Here we evaluate whether elementary modes can provide a new insight into the cause of ammonium toxicity.

In Figure 3.15 we plot the use efficiency of various elements with respect to the ammonia:nitrate uptake ratio of EFMs. We see that changes in the form of nitrogen assimilation are intimately coupled to widespread changes in potential steady state metabolism. This corresponds well to previous observation of widespread metabolic changes in response to the form in which nitrogen is supplied [158, 138].

It has previously been shown that the C:N ratio is greater in plants which can only utilise nitrate and not a mix of nitrate and ammonium [138]. Although we do not show that this must be true by the distribution of available elementary modes, in that this relationship does not hold for every mode, we do see that it is likely, assuming a somewhat parsimonious distribution of true biological fluxes amongst possible elementary modes. Interestingly, this occurs even without explicit modification of the biomass equation.

Arnold et al. [6] recently explained the preferentially mixed use of nitrate and ammonium in terms of energy efficiency, by showing that in their model, flux balance analysis predicted that mixed uptake of nitrate and ammonium allowed the most energy efficient production of some amino acids. This is because the fixed ratio of ATP:NADPH produced during photosynthesis and glucose catabolism could be altered, by using NADPH to reduce nitrate. Meaning that metabolism can access more energy efficient flux distributions, which were previously not usable due to their steady state ATP:NADPH requirements.

Here we see an alternative possible justification for mixed uptake. Figure 3.15 confirms that the most efficient metabolic forms in terms of energy, or nutrients required to produce biomass do use ammonium exclusively. This is presumably due to the smaller requirement for the production of reducing potential. However, these best case differences are extremely small in comparison to the magnitude of the worst case, least efficient elementary modes. Elementary modes which utilise exclusively either nitrate, or ammonium, can potentially be much less energy and nutrient efficient than modes which utilise a mix of the two. The least inefficient elementary modes for a given uptake ratio tend to be located in the region of 50:50 nitrate to ammonium uptake. If we parsimoniously assume an approximately equal distribution of flux between all elementary modes accessible at a particular ratio, the average efficiency is much greater for intermediate uptake than for the exclusive uptake of ammonium.

This suggests that plants may constrain available metabolism by controlling flux through ammonium and nitrate uptake transporters as an efficient means to restrict themselves to the 'least inefficient' regions of flux space, rather than regions where thy can potentially access the most efficient modes (although by

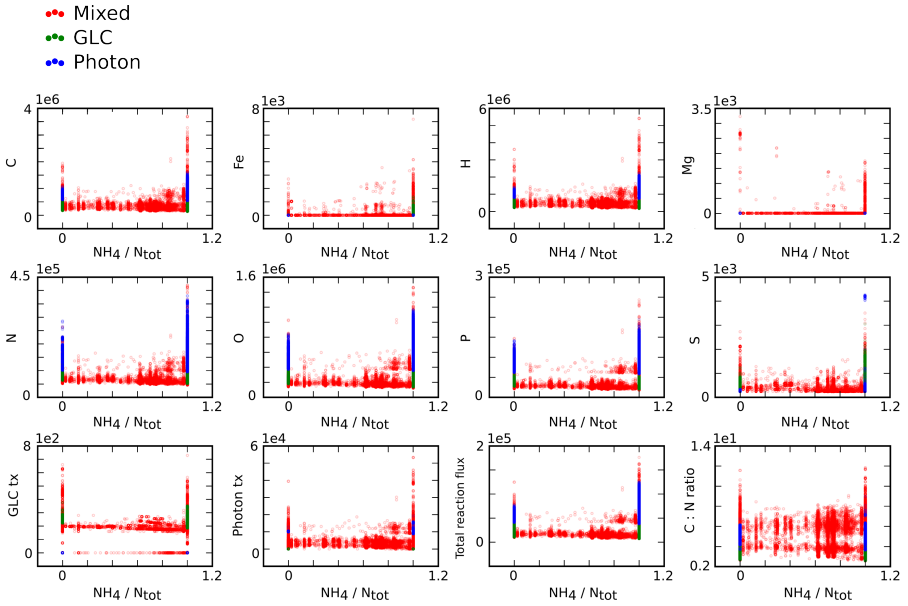



Figure 3.15: Fraction of N from $NH_4$ against other nutrient requirements. Each point corresponds to an elementary flux mode. In the legend, 'Photon' corresponds to EFMs in which uptake of glucose is not permitted, 'GLC' corresponds to flux modes in uptake of photons is not permitted, in 'mixed' flux uptake of both energy sources is allowed. Each horizontal axis is $NH_4$ / $N_{tot}$, such that at '0', all N is taken up as $NO_3$, at '1', all N is taken up as $NH_4$, at '0.5', both are taken up equally. In the top two rows, vertical axes indicate the element flux requirement of each EFM, (such that larger values are less nutrient efficient), in the bottom row, vertical axes correspond to flux through 'energy' uptake reaction, total internal flux, and the ratio of carbon to nitrogen required. Nutrient efficiency space available to the plant is described by the convex hull of the plotted points. In most cases, EFM distributions approximate a 'U' shape, indicating that modes which do not use an approximately equal $NH_4$: $NO_3$ ratio are on average less nutrient efficient than those which do. This suggests that nutrient use efficiency may explain the experimentally observed N use ratio.

only a small margin) but risk being extremely inefficient. This control via the preferential regulation of reactions which themselves exert strong steady state regulation of a large number of reactions has previously shown to be used in *E. coli* [90], and may be an efficient means of regulation, in terms of energy and resource investment, in comparison to maintaining a complex regulatory apparatus. Furthermore, the focus on limiting metabolic space to the 'least bad', rather than 'most efficient' regions through regulation is interestingly reminiscent of successful strategies for the rational design, using elementary modes, of efficient metabolite producing cell 'factories', in which genetic interventions are prioritised, based upon preventing flux through inefficient pathways, rather than increasingly the performance of the best ones [224].

This idea suggests that perhaps one reason for ammonium toxicity is that it permits more metabolic flexibility, and this allows access to on average less efficient elementary modes in terms of a large number of other resources. This does not imply that this is the only cause of toxicity, in what is likely a multi-pronged problem. However, the other barbs are likely to include the role of nitrate as a signalling molecule, and antioxidant, and the potential of ammonium to alter compartmental pHs, and are therefore difficult to address through constraint-based analysis directly.

### 3.3.6 Reactions controlling model behaviour

Having considered relationships between nutrient requirements, it is a natural step to also consider correlation between flux through particular reactions, and nutrient use efficiencies. This has two applications. Firstly to attempt to identify key diagnostic reactions which should be measured in order to understand which nutrients are being used. This is not necessarily easy to measure directly by assaying uptake from the environment, because plants can dissociate nutrient uptake from use, and use stores in the vacuole to buffer environmental exchanges [86]. Secondly, in identifying the reactions which most strongly correlate to a function of interest (in this case nutrient use), we identify potential targets for intervention in order to most strongly influence nutrient use efficiencies possible in metabolism in the plant. This is the logic behind 'Flux Design' [140], which in contrast to other EFM derived engineering predictions, allows identification of reactions for over-, as well as under-expression.

To initially evaluate the potential use of this approach, we plot correlation coefficients between reaction flux, and nutrient use efficiency for each reaction for each phenotype of interest. (Figure 3.16) shows that indeed, a relatively small number of reactions are highly correlated to each nutrient, and that the large majority of reactions are uninformative. This means that identifying correlated reactions is potentially interesting, as clearly reactions are informative to different degrees, and a small, comprehensible, number of reactions appear to be highly predictive of each behaviour of interest.

The identities of the identified 'informative' reactions must be carefully considered; if only 'obvious' reactions are recovered, then this method is not contributing anything of worth. Indeed, we do see that some returned reactions are trivial, for example flux through the $NH_4$ tx and $NO_3$ tx reactions are among the strongest predictors of the $NH_4$:$NO_3$ uptake ratio, however we also see interesting reactions as well. Table 3.6 shows the twenty reactions with the greatest correlation to elemental sulfur flux. Strikingly, these are generally more associated with the metabolism of active oxygen species than sulfur directly. The important sulfur containing metabolite glutathione is well known to be involved in plant response to reactive oxygen species stress (see [216] for a thorough review of plant sulfur metabolism), but it is not intuitive that these reactions would be better predictors for sulfur use than for example those involved in the

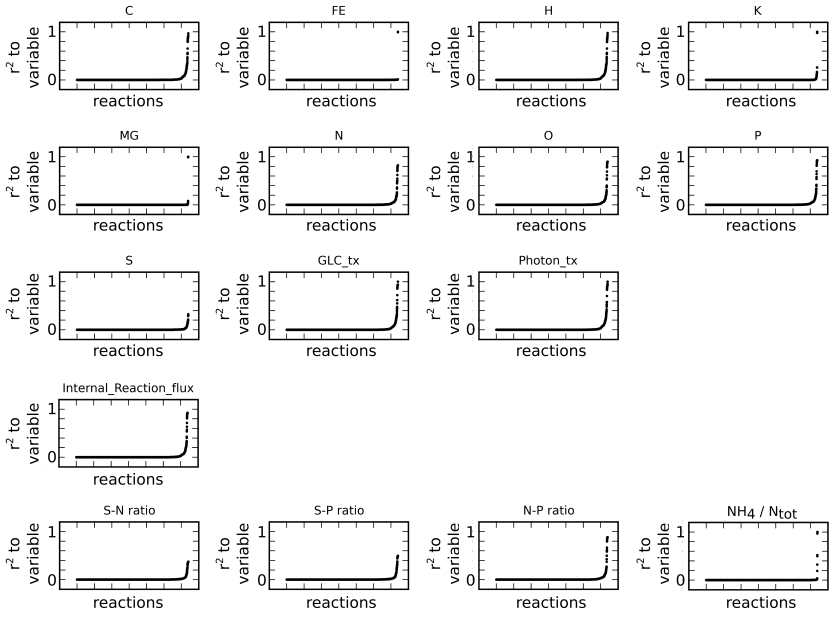



Figure 3.16: Variance of nutrient requirements explained by reactions. Each point is the coefficient of determination between flux through an individual reaction, and the criterion variable indicated in the facet. For each nutrient use variable assessed, the large majority of genes have only very small correlation to the variable, and a small set of reactions has strong correlation, suggesting that identifying the reactions which correlate to nutrient use phenotypes is a worthwhile exercise.

reduction of sulfate, or the formation of cysteine, or methionine. It therefore seems that the modelling approach taken does provide non-obvious insight into important reactions controlling particular phenotypes, and therefore identify potential targets for genetic engineering.

Due to the strong correlation between element flux, and total flux for many elements, we see that many reactions are likely returned primarily because of their influence on total flux through the elementary mode. For example Table 3.8 shows that many of the strongly correlated reactions are involved in central metabolism and in determining the balance between heterotrophy and autotrophy, (which we saw previously is indicative of total flux), rather than of nitrogen metabolism directly. Further supporting this, many of the nutrients strongly correlated to nitrogen use in Figure 3.14 also have many of the same highly correlated reactions as nitrogen use.

Ideally, any modification to metabolism should be as specific as possible. It is therefore ideal to pick reactions for modification which not only control the variable of interest, but also do not modify other variables. Using this correlation based approach, it is possible to evaluate the distribution of control between reactions for the different nutrient requirements, in order to determine firstly; whether such reactions exist, and secondly; to identify them. Figure 3.17 considers the example in which flux through the photon uptake reaction is desired to be modified. It shows that no reaction exists which can be expected to alter photon uptake, without also strongly effecting nitrogen elemental-flux, due to the tight grouping of all points to the best fit line. However reactions can be seen to exist which can be targeted with strong control coefficients over photon flux, but less over phosphorus, or sulfur requirements.

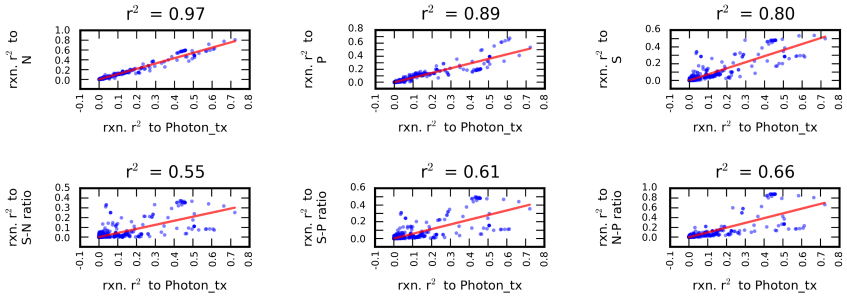



Figure 3.17: Correlation of explanatory power of reactions between photon uptake, and other elemental nutrient flux. Many of the same reactions correlate strongly to multiple nutrient phenotypes of interest, however the degree of correlation varies. This has implications for target selection in order to manage the specificity of any genetic interventions, and is likely a consequence of correlation between many phenotypes, and total flux.

Interestingly in all cases assessed, correlation between reaction control of the phenotypes of interest is strongly positive, suggesting that generally reactions

Table 3.6: The twenty reactions with the largest $R^2$ to total elemental S flux. Interestingly, we see that the reactions are not necessarily the sulfur uptake and reduction genes which might be expected. We see a number of reactions associated with reactive oxygen species, presumably due to the role of sulfur containing metabolite glutathione in managing the accumulation of these chemicals [4].

| Reaction ID | $R^2$ | Reaction formula | Associated pathways |
|---|---|---|---|
| SUCCINATE-DEHYDROGENASE--UBIQUINONE-RXN m | 0.32 | succinate [m] + ubiquinone [m] ⟹ fumarate [m] + ubiquinol [m] | aerobic-respiration-III aerobic-respiration-I TCA-cycle-II |
| GLUTATHIONE-REDUCT-NADPH-RXN m | 0.3 | NADPH [m] + oxidized-glutathione [m] ⟹ 2 glutathione [m] + NADP [m] | glutathione-glutaredoxin redox reactions glutathione-peroxide redox reactions |
| 1.8.5.1-RXN m | 0.3 | 2 glutathione [m] + L-dehydro-ascorbate [m] ⟹ ascorbate [m] + oxidized-glutathione [m] | ascorbate glutathione cycle |
| L-ASCORBATE-PEROXIDASE-RXN m | 0.29 | ascorbate [m] + hydrogen peroxide[m] ⟹ L-dehydro-ascorbate [m] | |
| MALSYN-RXN x | 0.21 | acetyl-CoA [x] + glyoxylate [x] ⟹ coenzyme-A [x] + malate [x] | glyoxylate-bypass glycolate and glyoxylate degradation II |
| CO2 mc_f | 0.21 | carbon-dioxide [m] ⟹ carbon-dioxide [c] | |
| SUPER-OXIDE-GEN[m] | 0.2 | ⟹ super-oxide [m] | |
| SUPEROX-DISMUT-RXN m | 0.2 | 2 super-oxide [m] ⟹ hydrogen peroxide[m] + oxygen [m] | reactive oxygen species degradation |
| ISOCITDEH-RXN m_f | 0.2 | NADP [m] + D-threo-isocitrate [m] ⟹ 2-ketoglutarate [m] + carbon-dioxide [m] + NADPH [m] | L-glutamine biosynthesis-III |
| FUMHYDR-RXN m_f | 0.19 | fumarate [m ⟹ malate [m] | PWYQT-4481 PWY-5690 |
| THREO-DS-ISO-CITRATE xc_b | 0.17 | D-threo-isocitrate [c] ⟹ D-threo-isocitrate [x] | |
| SUC xc_f | 0.17 | succinate [x] ⟹ succinate [c] | |
| ISOCIT-CLEAV-RXN x | 0.17 | D-threo-isocitrate [x] ⟹ glyoxylate [x] + succinate [x] | glyoxylate-bypass |
| 2OXOGLUTARATEDEH-RXN m | 0.17 | 2-ketoglutarate [m] + coenzyme-A [m] + NAD [m] ⟹ carbon-dioxide [m] + NADH [m] + succinate-CoA [m] | TCA-cycle-II |
| Pi xc_f | 0.15 | phosphate [x] ⟹ phosphate [c] | |
| AMP ATP xc_f | 0.15 | AMP [x] + ATP [c] ⟹ ATP [x] + AMP [c] | |
| INORGPYROPHOSPHAT-RXN x | 0.15 | diphosphate [x] ⟹ 2 phosphate [x] | |
| ACET xc_b | 0.15 | acetate [c] ⟹ acetate [x] | |
| ACETATE–COA-LIGASE-RXN x | 0.15 | acetate [x] + ATP [x] + coenzyme-A [x] ⟹ acetyl-CoA [x] + AMP [x] + diphosphate [x] | salicylate glucosides biosynthesis I |
| MAL xc_f | 0.15 | malate [x] ⟹ malate [c] | |

Table 3.8: The twenty reactions with the largest $R^2$ to total elemental N flux, appear to be unrelated to N specific metabolism, but related to autotrophy versus heterotrophy, and total flux. This is also true of many other nutrient phenotypes assessed, and is likely a consequence of the common correlation between nutrient efficiency, and total flux.

| **Reaction ID** | $R^2$ | **Reaction formula** | **Associated pathways** |
|---|---|---|---|
| CO2 pc_b | 0.82 | carbon-dioxide [c] ⟹ carbon-dioxide [p] | |
| RIBULP3EPIM-RXN p_b | 0.82 | xylulose-5-phosphate [p] ⟹ ribulose-5-phosphate [p] | calvin pathway rubisco shunt pentose phosphate pathway (non-oxidative branch) |
| RIB5PISOM-RXN p_f | 0.82 | ribose-5-phosphate [p] ⟹ ribulose-5-phosphate [p] | calvin pathway rubisco shunt pentose phosphate pathway (non-oxidative branch) |
| RIBULOSE-BISPHOSPHATE-CARBOXYLASE-RXN p | 0.82 | carbon-dioxide [p] + D-ribulose-1_5-bisphosphate [p] ⟹ 2 3-phospho-D-glycerate [p] | calvin pathway rubisco shunt |
| PHOSPHORIBULOKINASE-RXN p | 0.82 | ATP [p] + ribulose-5-phosphate [p] ⟹ ADP [p] + D-ribulose-1_5-bisphosphate [p] | calvin pathway rubisco shunt |
| 1TRANSKETO-RXN p_f | 0.82 | D-sedoheptulose 7-phosphate [p] + D-glyceraldehyde 3-phosphate [p] ⟹ ribose-5-phosphate [p] + xylulose-5-phosphate [p] | calvin pathway rubisco shunt pentose phosphate pathway (non-oxidative branch) |
| 2TRANSKETO-RXN p_b | 0.82 | fructose-6-phosphate [p] + D-glyceraldehyde 3-phosphate [p] ⟹ erythrose-4-phosphate [p] + xylulose-5-phosphate [p] | calvin pathway rubisco shunt pentose phosphate pathway (non-oxidative branch) |
| O2 pc_f | 0.81 | oxygen [p] ⟹ oxygen [c] | |
| CO2 tx_f | 0.81 | ⟹ carbon-dioxide [c] | |
| O2 tx_b | 0.81 | oxygen [c] ⟹ | |
| 1.18.1.2-RXN p_f | 0.81 | NADP [p] + 2 Reduced-ferredoxins [p] ⟹ NADPH [p] + 2 Oxidized-ferredoxins [p] | glycolysis IV |
| Plastidial ATP Synthase p | 0.79 | 3 ADP [p] + 3 Pi [p] + 4 Pumped-proton [p] ⟹ 3 ATP [p] | |
| RXN490-3650 p | 0.78 | Oxidized-ferredoxins [p] + Photon [p] + Plastocyanin-Reduced [p] ⟹ Oxidized-Plastocyanins [p] + Reduced-ferredoxins [p] | photosynthesis light reactions |
| PLASTOQUINOL-PLASTOCYANIN-REDUCTASE-RXN p | 0.78 | 2 Oxidized-Plastocyanins [p] + plastoquinol [p] ⟹ plastoquinone [p] + 2 Plastocyanin-Reduced [p] + 4 Pumped-proton [p] | photosynthesis light reactions |
| Photon tx | 0.75 | ⟹ Photon [p] | |
| PHOSGLYPHOS-RXN p_f | 0.72 | ATP [p] + 3-phospho-D-glycerate [p] ⟹ ADP [p] + 1_3-bisphospho-D-glycerate [p] | glycolysis IV sucrose biosynthesis I glycolysis calvin pathway gluconeogenesis glycolysis II gluconeogenesis III |
| TRIOSEPISOMERIZATION-RXN p_f | 0.72 | D-glyceraldehyde 3-phosphate [p] ⟹ dihydroxy-acetone-phosphate [p] | glycolysis IV glycolysis calvin pathway glycolysis II gluconeogenesis III |
| GLC tx | 0.7 | Pumped-proton [c] ⟹ glucose [c] | |
| PSII-RXN p | 0.62 | 2 plastoquinone [p] + 4 Photon [p] ⟹ oxygen [p] + 2 plastoquinol [p] + 4 Pumped-proton [p] | photosynthesis light reactions |
| H+ ATPase c | 0.55 | ATP [c] ⟹ ADP [c] + Pi [c] + Pumped-proton [c] | |

are important, or not important across multiple phenotypes, and that therefore unless this step is performed, interventions can be expected to have severe off target effects. This suggests an extension to the 'Flux Design' [140] approach in which rather than prioritisation based on $r^2$ to the phenotype of interest alone, reactions are prioritised for modification based on a score

$$s = \frac{r_i^2}{\frac{1}{m}\sum_{j=1}^{m} c_j r_j^2}, \tag{3.1}$$

in which $r_i^2$ is the explanatory power of the reaction to the reaction to the phenotype flux of interest, $\vec{c} \in \mathbb{R}^{\mathrm{m}}$, is a vector of weights, prioritising the maintenance of phenotypes, and $r_j^2$ is the explanatory power of the reaction for phenotype $j$ which is desired to be unaffected, and $m$ is the number of phenotypes considered, where phenotypes are considered to be functions of reaction sets.

Inspired by Behre et al. [16] we extend this metric to the consideration of multiple knockouts, such that rather than $r_i^2$, and $r_j^2$ corresponding to the squared correlation between reaction flux, and phenotype flux, it corresponds to the R-squared between phenotype flux, and a regression model, in which $i$ reaction fluxes are predictors. This allows the quantification of the suitability of a given phenotype target as

$$\frac{1}{R}\sum_{i=1}^{R} \max(s_i), \tag{3.2}$$

where $R$ is the number of reactions in the model, and $s_i$ is the reaction score for a set of $i$ knockout reactions defined in Equation 3.1. Although obviously given the explosion in the number of reaction knockout combinations which must be considered, we expect that, similar to the metrics proposed by Behre et al. [16], this value is not practically calculable, it is likely to be reasonably approximated by considering only small reaction knockout sets, and in particular can be calculated for the relatively small number of interventions experimentally possible.

Generally, we observe that it seems more possible to tweak relative nutrient requirements (bottom row of Figure 3.17), without influencing other variables. For example it is more possible to control the N:P requirement ratio without modifying photon uptake, than it is to modify nitrogen, or phosphorus requirement alone. It will be interesting to apply some approximate form of Equation 3.2 in order to explore whether particular regions of metabolism are more suitable for engineering than others.

The apparent interrelatedness of control coefficients for many nutrient requirements indicates a control structure in which reactions typically either have strong influence over many nutrient variables, or none. This is consistent with hierarchical flux coupling studies which have found that in many organisms reactions often control a subset of subsidiary reaction, but that this control is not necessarily symmetrical, these key driver reactions are themselves targets for stringent regulation by the organism [90]. Although this means that

it is predicted that modification of these reactions will lead to the most dramatic changes in metabolism, they do not allow precise control of individual phenotypes. Deciding the appropriate tradeoff between control coefficients and ubiquity of effect is likely to be situation dependent, but this correlation based approach provides a framework through which the relative merits can at least be assessed.

### 3.3.7 Conclusion

We have seen that through the unbiased consideration of the elementary modes of a metabolic network of Arabidopsis, rather than the 'optimal' solutions generated by flux balance analysis type approaches, we are able to gain insights into metabolism which are not otherwise possible. This was exemplified most clearly in proposing an explanation for the observed nitrogen uptake ratio of ammonia to nitrate which explicitly depends upon restricting the distribution of suboptimal flux distributions to the 'least bad' nutrient efficiency region at the experimentally observed value, an idea which would not be possible through FBA.

By studying the capabilities of a metabolic network, we also demonstrated the limitations of a conventional 'uptake flux' based assessment of metabolic efficiency when applied to chemical substrates, due to the extreme inflexibility in detectable efficiency, and developed an alternative metric. This also highlighted the importance of permitted 'waste' metabolites in genome scale models, which seems to be generally overlooked. Using the element flux metric, we found a surprising inflexibility in nutrient requirements, and the ability to trade-off requirements against each other through the selection of alternative flux distributions. This suggests that in contrast to robustness to genetic perturbation, perhaps alteration to the biomass equation may be a more important component of environmental robustness than the structure of the reaction network. We have also explored the use of elementary modes for prioritising targets for genetic engineering, and proposed extensions to existing methods. We emphasise that none of these developments would be possible through alternative constraint-based analytical frameworks.

A common theme, however, has been the problem that the EFMs assessed are potentially too unbiased. Flux balance analysis suffers from the imposition of arbitrary optimality criteria, which potentially restrict the recovered solutions too much, or to incorrect regions. However, the point of biological regulation of metabolism is to restrict flux space to some subset of the steady state space described by elementary modes, and this idea is lost in the analysis described in this chapter. Metabolic regulation means that it is unlikely that all possible elementary modes are used by the plant, and it is often unclear how applicable the behaviour of the full EFM set is to the behaviour of those which are actually biologically utilised. This shortcoming has been clear, for example when considering nutrient tradeoff requirements, (Figure 3.14). Although in the whole set of

EFMs defined by the metabolic network strong correlations may exist, between for example nitrogen and phosphorus requirements, it is not clear how well this full set of EFMs approximates the subset which Arabidopsis metabolism is able to actually use, and therefore how much confidence there can be in any found nutrient relationships.

The tradeoff between imposing arbitrary assumptions to restrict considered steady state flux solution space, and not being able to make predictions is currently difficult to resolve. It is interesting that flux balance analysis variations have recently been developed which attempt to sample from the solution space, whilst still imposing optimality criteria [32]. Furthermore, a number of methods attempt to integrate additional experimental data to restrict, or weight, the set of elementary modes considered [239, 153, 236]. It seems likely that these two approaches will converge somewhere in the middle, although it is currently unclear whether somewhat arbitrary FBA optimality criteria, or EFMs constrained using essentially inadequate, but widely available transcriptomic data will yield the more accurate descriptions of metabolism.

## 3.4 Methods

### 3.4.1 TreeEFM

TreeEFM [163], is an algorithm for the calculation of EFMs, and at the time of writing is among the best performing of such tools. Here I will briefly summarise the approach. As previously discussed in chapter 2, a metabolic network consisting of $C$ metabolites, and $R$ reactions can be represented by the stoichiometric matrix $S$, where element $S_{\{c,r\}}$ is the stoichiometric coefficient for metabolite $c$ in reaction $r$. By convention, substrates are assigned negative coefficients, and products positive. For each reaction $r$ in $(r = 1, ..., R)$ a continuous variable, $v_r$, represents flux through it. Steady state reaction sets can be found by solving the equation

$$\sum_{r=1}^{R} S_{\{c,r\}} v_r = 0, \quad \forall c \in I \tag{3.3}$$

where $I$ represents the internal metabolites of the metabolic model. EFMs are the vertices of the 'flux cone' defined by this equation. By cutting this cone with a hyperplane $v_r = 1$, a projection of the flux cone is generated, the extreme points of which correspond to EFMs. In the work presented here the reaction $r$ was chosen to be the so-called 'biomass equation', previously defined in chapter 2, in which all metabolites considered to be essential for the production of biomass in Arabidopsis are produced, and exported from the model.

Efficient methods for the enumeration of extreme points have been previously developed, in the particular implementation of the TreeEFM algorithm used

here, the Simplex algorithm [45] was used.

In order to calculate diverse subsets of EFMs, a tree based approach is used to recursively modify the linear programming problem. Each node of the tree represents a linear program with a different solution set to its ancestors. If the elementary flux mode solution to an arbitrary parent node, consists of $m$ reactions with non-zero flux, for each non-zero reaction r ($r = 1, ..., m$) a daughter node is generated with the additional constraint $v_r = 0$.

This is the basis of the method, although in the implementation provided by Pey et al. [163] and used here, additional heuristics are used to prioritise the solution of particular nodes, based on the number of previously found EFMs which are also feasible in the node.

EFMs are returned in the form of a matrix, $E$. where each row corresponds to an elementary flux mode, and each column corresponds to a reaction. Each element, $E_{\{e,r\}}$, of the matrix corresponds to flux through reaction $r$ in elementary flux mode $e$, normalised by flux through the biomass reaction.

## 3.4.2 Databases

### 3.4.2.1 Aracyc

The Aracyc database ([253], version 13.0) is one of 7600 pathway/genome databases available as part of the biocyc database collection (`biocyc.org`). It is considered a "Tier 1" database, meaning that it was "created through intensive manual efforts", and is "constantly updated". It can therefore be considered a high quality source of information about the metabolic reactions possible within the organism *Arabidopsis thaliana*, and to offer a good representation of current biological thinking about which pathway(s) a reaction is considered a part of. This dataset also includes information about the position of reactions within a pathway.

### 3.4.2.2 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG, [50], release 77.0, `www.kegg.jp/`) is a database of the relationships of a number of biological entities. 'KEGG-PATHWAY' was used in this chapter as a second manually curated database the the reactions found within metabolic paths.

### 3.4.2.3 Expression Atlas

Expression Atlas ([161], `www.ebi.ac.uk/gxa/home`) provides a database of gene expression patterns under different biological conditions. A standardised 'in-house' data processing and analysis pipeline is carried out on all raw data to

allow comparisons between and across experiments. Here data from 6,719 microarray chip experiments across 462 published studies of Arabidopsis was used to calculate the correlation coefficients between all pairs of genes which could be associated to the Aracyc or KEGG reaction identifiers.

As there is not necessarily a 1:1 relationship between a gene product and a reaction, the Expression Atlas was preprocessed such that in instances in which more than one gene product mapped to a reaction, the mean gene product value was used. When more than one reaction mapped to a single gene product, the transcript was associated with all mapped reactions. Consequently, as the Aracyc and KEGG databases do not differentiate between the same reaction in different compartments, genes were mapped to all instances of a reaction across compartments.

### 3.4.3 Reaction Clustering

#### 3.4.3.1 Reaction distance

Within the Aracyc and KEGG databases, each reaction is either present or absent from a given internally defined biological pathway. This membership can be represented in each case by the binary membership matrix $M$ consisting of $R$ reactions, and $P$ paths, where $M_{\{r,p\}}$ is 1 if reaction $r$ is considered to be in path $p$, and zero otherwise.

The similarity between reactions $r$, and $s$ was calculated as the number of pathways in which both reactions were involved, (i.e. the number of positions for which $M_{\{r,i\}} + M_{\{s,i\}} = 2$). This was normalised by the number of paths in which either reaction were involved, (i.e. the number of positions for which $M_{\{r,i\}} + M_{\{s,i\}} \neq 0$), so as to account for different reaction promiscuity, and map similarities onto to a [0,1] interval. The distance between two reactions was calculated as 1 minus the similarity.

We considered joint membership of all pathways in the database to carry an equal weight , because it is not clear how any non-equal cost weighting should be assigned. We normalised only by the number of non-zero entries, rather than all entries, because most reactions are present in only a small number of pathways, and normalising by all entires (for which the majority which involve neither reaction $r$, nor $s$), would compress the differences between reaction pairs.

Reaction distance in the Expression Atlas, and EFMs datasets were calculated as

$$d_{r,s} = \sqrt{2 \cdot (1 - C_{r,s})} \tag{3.4}$$

where $C_{r,s}$ is the correlation coefficient between reactions $r$ and $s$ either in terms of the abundance of their associated transcript(s) across different experiments, or their flux in different EFMs respectively.

Correlation coefficients were calculated either directly from flux through calculated elementary modes which are able to produce biomass, or via null-space analysis as described by Poolman et al. [170]. $\phi_{uv}$, the cosine of the angle between rows $u$ and $v$ of the kernel matrix, K

$$\phi_{uv} = \frac{\vec{u} \cdot \vec{v}}{||\vec{u}|| \cdot ||\vec{v}||} \tag{3.5}$$

can be shown to be mathematically equivalent to the Pearson's correlation coefficient between the reactions $u$ and $v$ across all EFMs [170].

#### 3.4.3.2 Clustering

Hierarchical clustering of reactions was performed within each dataset using distance matrices calculated as described above and functions in the *scipy.cluster.hierarchy* module (v0.17.1). The 'average' linkage method was used, such that the distance between clusters $u$, and $v$ was calculated as

$$D(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{|u| * |v|} \tag{3.6}$$

where $i$ and $j$ iterate over all reactions in the clusters $u$, and $v$, respectively, and $d$ is the function defined in either paragraph two of subsubsection 3.4.3.1, or Equation 3.4.

#### 3.4.3.3 Cluster similarity

For comparison of the reaction clusters produced using different datasets, an initial preprocessing step was included such that only reactions which could be mapped to both datasets were considered.

For each reaction cluster, $u$, produced at a given cuttoff linkage distance in the 'test' dataset, the similarity measure $F_1$ was calculated as

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.7}$$

to each of the clusters, $v$, generated from the other 'comparison' dataset. For each cluster in the 'test' dataset, the $F_1$ to the most similar cluster in the 'comparison' dataset is reported.

Precision is the shared number of reactions between the compared clusters, i.e. the 'correct' reactions in cluster $v$, as a fraction of the total number of reactions

in $v$,

$$precision = \frac{|u \cap v|}{|v|}, \tag{3.8}$$

and therefore penalises larger, less precise $v$ groups. Recall is the fraction of the reactions in the group $u$, which are also in the compared group $v$,

$$recall = \frac{|u \cap v|}{|u|}, \tag{3.9}$$

and therefore penalises the failure to recover all of the expected reactions.

#### 3.4.3.4 Gene ontology annotation enrichment within clusters

For each cluster, each reaction was mapped to gene ontology annotations via associated genes. Each cluster was scored using

$$s = 1 - \frac{unique}{total}, \tag{3.10}$$

where *unique* is the number of unique gene ontology terms associated with reactions in the cluster, and *total* is the number of gene ontology annotations associated with any reactions in the cluster, including duplicate annotations associated with multiple reactions.

The score for each set of clusters at a given linkage threshold was calculated as the mean score of all clusters at that distance. To estimate the probability achieving of each score by chance, random clusters of reaction identities were assigned with the same cluster size distribution as the true set. The fraction of 10,000 such random assignations which achieved the same, or better score that the true clustering was calculated.

#### 3.4.3.5 The effect of gene ontology on reaction clustering performance

For each gene ontology annotation, all associated genes were mapped to reactions in the modified Arabidopsis model described in chapter 2. The $F_1$ score for the annotation was calculated as the mean $F_1$ of the clusters for which these reactions were members.

The probability of achieving a given gene ontology $F_1$ score by chance was numerically approximated by randomly sampling an equivalent number of reactions from the 'test' dataset 100,000 times, and finding the proportion of random reaction sets with an equal or better mean $F_1$ score.

The Bonferroni correction was applied, in which each approximated probability is multiplied by the number of tests carried out. In this case, this is the number of different gene ontology annotations associated with reaction in the model.

### 3.4.4 Nutrient requirement metrics

Two methods for deriving nutrient requirement phenotype predictions from EFMs are discussed. In the first, flux ($v_r$) through nutrient uptake reactions are normalised by flux through the biomass equation in order to directly assess nutrient use efficiency. Total flux is calculated as the sum of each row of the elementary flux mode matrix $E$.

In the second method, elemental flux, matrix $T$ is used, in which rows correspond to reactions, and each column corresponds to one of $C$ chemical elements, $T_{r,c}$ is the stoichiometric coefficient of element $c$ in reaction $r$, following the convention that substrates are negative, and products positive. $T$ is therefore analogous to the transpose of the stoichiometric matrix $S$, with metabolites replaced by elements.

As the model is mass balanced, such that elements are neither created, nor destroyed by internal reactions, the amount of each element in flux through a reaction can be equivalently expressed in terms of reaction flux multiplied by either substrate stoichiometry, or product stoichiometry.

To prevent 'double-counting' through the addition of substrate and product flux for each reaction, $T^p$, is the transformation of matrix T, such that all negative values (i.e. substrates) are replaced by zero. Elemental flux therefore is given by

$$F = E \cdot T^p \tag{3.11}$$

where $F_{e,c}$ corresponds to total flux of chemical element $c$ in elementary flux mode $e$, normalised by flux through the biomass reaction. 'Total elemental flux' is the sum of each row of matrix F.

### 3.4.5 Predictor reactions

In order to determine the reactions which are directly or indirectly dominant in determining the emergent behaviour of the modelled system in terms of nutrient use requirements, the *scikit-learn* (v0.17.1) toolkit was used to learn all linear regression models, and calculate the coefficients of determination ($r^2$) values. To asses one and two parameter linear models, a brute force approach was used to enumerate all combinations of one or two indicator reactions for each criterion variable.

# Chapter 4

# The regulation of mobile mRNA

Recently, a large population of messenger RNA (mRNA) was shown to be able to travel between plant organs via sieve elements as a putative long-distance signalling molecule. However, a mechanistic basis by which transcripts are selected for transport has not yet been identified. Here we show that experimental mRNA mobility data in Arabidopsis can be explained by transcript abundance and half-life. This suggests that the majority of identified mobile transcripts can be accounted for by non-sequence-specific movement of mRNA from companion cells into sieve elements.

## 4.1 Introduction

Acclimation to environmental conditions is vital for plants. At the whole-plant level, this is aided by long distance signalling between organs, which is important both for plant development and defence responses [198, 208]. Mechanisms for long distance communication include calcium and ROS waves, action potentials, and hydraulic waves, as well as phytohormones and some small RNAs [198, 208, 67]. Long-distance signalling molecules can be transported through the phloem, in enucleated cells called sieve elements. mRNA is also able to move in sieve elements, and some mobile transcripts have been shown to give rise to developmental differences at distal locations [209], leading to the suggestion that mRNA could be another class of long distance signalling molecules [241].

mRNA moves between host and parasitic plants [107] as well as between heterografts [154]. Recently a pioneering grafting approach identified a large population of 2,006 mobile mRNA species that were able to move between roots and shoots in grafted Arabidopsis ecotypes [219]. Interestingly, these data suggest

that a large percentage of mRNA can move against the direction of phloem flow. Phenotypic changes related to specific mobile mRNAs have been reported [108, 84, 9, 133, 155] but it remains unclear to what extent mRNA mobility is biologically meaningful [129, 154]. Correlation between abundance and long distance mobility has been noted, leading to speculation that mRNA transport could occur in both a selective and non-selective manner [107, 154].

Here, we investigate the potential link between mRNA abundance and mobility by evaluating a simple diffusion-based model (termed the abundance model) in which nonsequence-specific movement of mRNA species from companion cells into sieve elements leads to long-distance mobility. We find that this model is sufficient to explain the large population of experimentally observed mobile transcripts, and makes predictions regarding mobile transcript size and half-life that are consistent with experimental data.

This analysis suggests that most of the identified mobile mRNA species are mobile as a consequence of local abundance.

## 4.2 Results

### 4.2.1 The probability of mRNA mobility saturates with mRNA abundance

We developed a simple model to estimate the probability that a transcript is mobile. In the companion cell, after mRNA crosses the nuclear envelope, most transcripts move through the cytosol by diffusion [60] and are translated or degraded. Alternatively, upon reaching the cell membrane, the transcripts may pass or be chaperoned through plasmodesmata into sieve elements [103, 154], in which molecules can move bi-directionally [126]. The fate of each mRNA molecule of a given transcript species was modelled by a random walk through a 3D cell. Initially positioned at the centre of the cell, at each time-point, the molecule could move up, down, left right, forward or back a small distance relative to the size of the cell, or it could decay with a predefined, constant probability. At the cell boundary, a spatially uniform, transcript independent probability that the mRNA could pass into sieve elements was assigned. If any simulated mRNA molecule passed through the cell membrane, then that transcript was considered to be mobile (Figure 4.1a). These assumptions could be readily extended to include further information, such as varying plasmodesmal densities, but whilst simple they proved to be sufficient to explain the observed data, thus not warranting further parameters in the model.

mRNA species abundance is a consequence of transcription rate and half-life. Different transcription rates were modelled by changing the initial number of mRNA molecules in the simulation, and half-life by the decay probability. Transport across the graft boundary in sieve elements was assumed to be fast, and

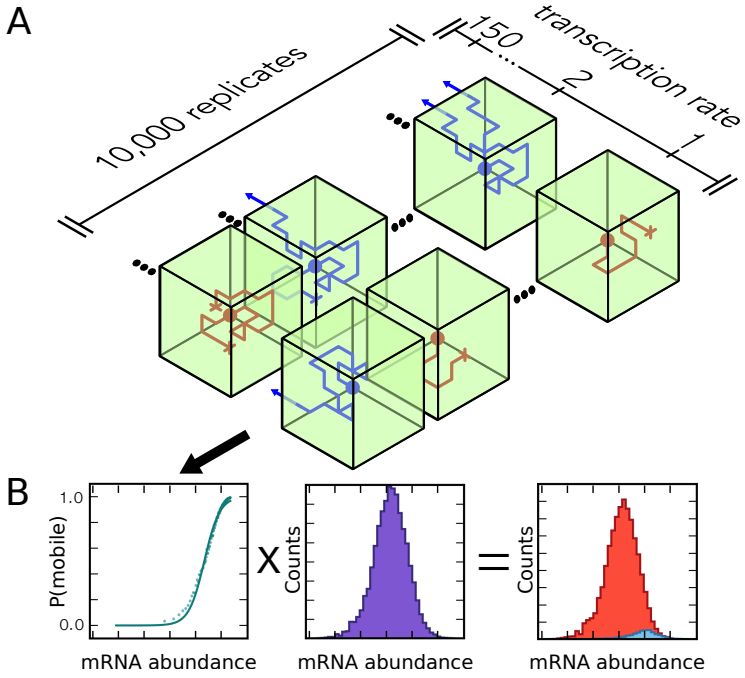



Figure 4.1: Workflow for the simulation of mRNA mobility. A, cartoon of the mRNA abundance model. Green boxes represent cells, the side facing inwards being adjacent to sieve elements. The rows of cells represent different simulation runs and different transcription rates. Blue diffusion paths indicate simulations in which the transcript was considered to be mobile, and red indicates those in which the transcript was non-mobile. B, the output of the abundance model (left, mobility versus abundance plot) was combined with experimental mRNA abundance data to predict the distributions for mobile and non-mobile mRNA (see Methods).

so if any molecule of a transcript passed into sieve elements, then that transcript was considered to be mobile, otherwise the transcript was considered non-mobile. Modelled mRNA species fate was seen to be stochastic, and so for each transcript species, the simulation was run 10,000 times to estimate a probability of movement out of the cell, which was then used to calculate the probability of an mRNA species moving into sieve elements from multiple companion cells.

This simple model predicts a saturation relationship between mRNA abundance and probability of mobility. The shape of this curve depends on a number of variables such as cell size, the number of companion cells, plasmodesmatal conductivity, and nucleus size and position, as well as mRNA half-life, but can be approximated by a saturation curve with only two unknown parameters (Figure 4.2, Figure 4.3, Figure 4.4, Methods).

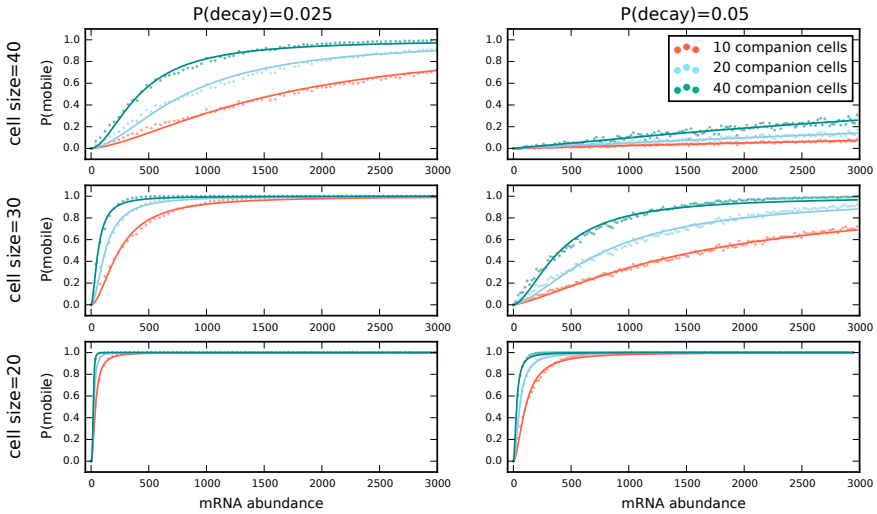



Figure 4.2: The effect of cell size, cell number and half-life on mRNA mobility. The mRNA abundance model predictions for the effect of species abundance on mRNA mobility, and the saturation curve approximations of these are shown for simulations with varied cell size, cell number, and mRNA half-life characterised by the probability of decay, P(decay). Points are model predictions based on the calculation of the escape probability from a companion cell in the 3D cell simulations described in the main text. The curves correspond to a fitted saturation equation for a transcript being mobile as explained in the Methods section. This shows that although these unknown parameters influence mobility, the effects can be well approximated by a simple saturation curve, thus reducing the number of unknown parameters in the model.

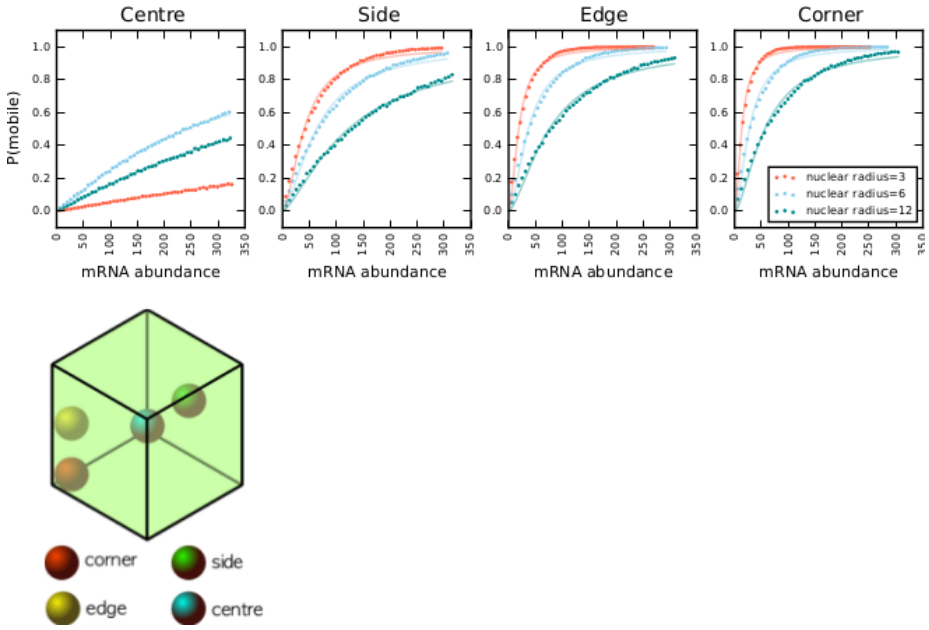



Figure 4.3: The effect of nucleus position and size on mRNA mobility. The abundance model predictions for the effect of species abundance on mRNA mobility and the saturation curve approximations of these are shown for simulations with varied nucleus position and size. Variation in nucleus size and position was incorporated in the abundance model by modifying the initial position of the simulated mRNA molecule. Starting positions were uniformly sampled on the surface of a sphere, representing the nucleus, positioned as shown, and of stated radius. Points are model predictions based on the calculation of the escape probability from a companion cell in the 3D cell simulations described in the main text. The curves correspond to a fitted saturation equation for a transcript being mobile as explained in the Methods section. This shows that although nucleus size and position influence mobility, the effects can be well approximated by a simple saturation curve, thus reducing the number of unknown parameters in the model.

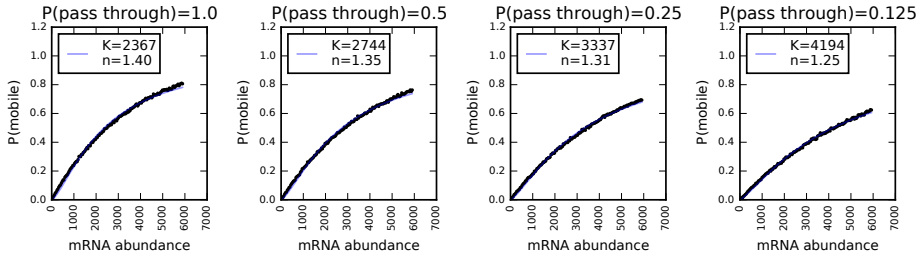



Figure 4.4: The effect of varied probability of passing through the cell surface on mRNA mobility. The abundance model predictions for the effect of species abundance on mRNA mobility and the saturation curve approximations of these are shown for simulations with varied probability of passing through the cell surface upon contact with it. With decreased probability of passing through the surface, $K$ was seen to increase, and $n$ decrease. Points are model predictions based on the calculation of the escape probability from a companion cell in the 3D cell simulations described in the main text. The curves correspond to a fitted saturation equation for a transcript being mobile as explained in the Methods section. This shows that although these unknown parameters influence mobility, the effects can be well approximated by a simple saturation curve, thus reducing the number of unknown parameters in the model.

### 4.2.2 The predicted abundance distribution of mobile transcripts fits experimental data

We compared the predicted relationship between mRNA abundance and mobility from the model to the dataset generated by Thieme et al. [219] (Figure 4.1b, Methods). With fitted parameters (Figure 4.5a, Methods), the computed relationship between transcript abundance and probability of mobility was able to reproduce the distribution of the mobile and non-mobile mRNA species (Figure 4.5b), although as expected, the fate of individual transcripts was highly stochastic. This was also observed within the experimental data, where transcripts frequently could be mobile or not in different repeats. As can be seen in Figure 4.5a & c, the predictions remained within experimental error; however, the experimental data seemed to deviate from the model at both extremes of the transcript abundance distribution. This is likely predominantly a consequence of the low copy statistics for mRNA species of extreme abundances (Figure 4.5b), although it could indicate the existence of an alternative mechanism affecting a small proportion of the population, which is otherwise hidden by the abundance-driven mobility mechanism.

### 4.2.3 Analysis of low-abundance mobile transcripts

To investigate whether there are differences in the nature of the transcripts that deviate most from our simple abundance model, we analysed the sequences of

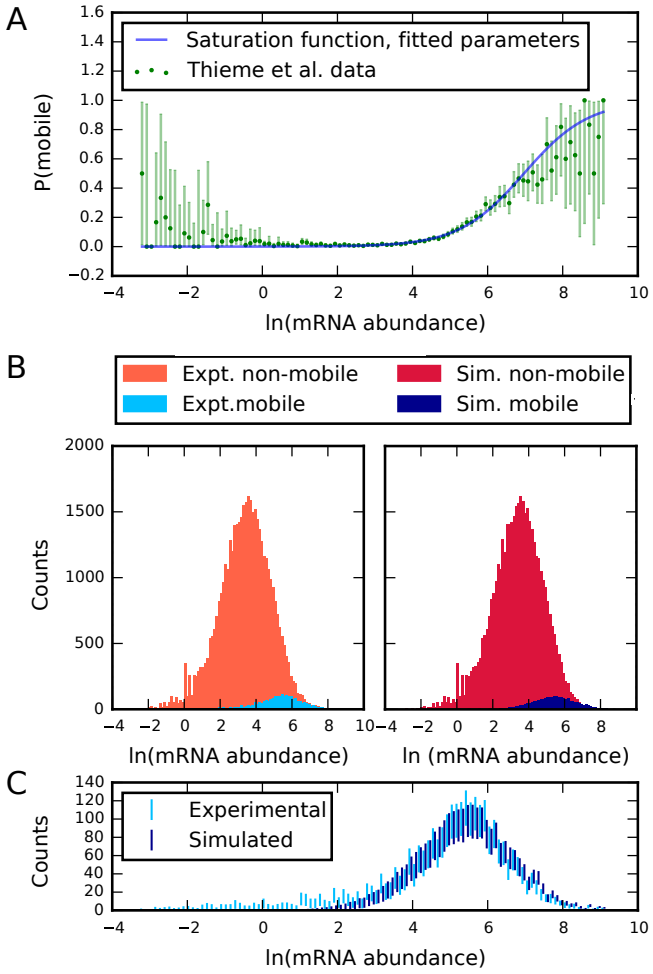



Figure 4.5: An abundance model of mRNA mobility captures the experimental mRNA distributions. A, the fitted and experimentally derived P(mobile) curve, 95% experimental confidence intervals shown. The experimental P(mobile) was estimated as the ratio of the number of mobile over non-mobile transcripts, binned by abundance. The confidence interval was calculated using the Clopper-Pearson Exact Binomial method. At high and low mRNA abundance values, experimental P(mobile) uncertainty increases as a consequence of the fewer number of high and low abundance transcript species. B, experimental (left), and modelled (right) abundance distributions of mobile and non-mobile mRNA using the fitted P(mobile) curve. C, experimental and predicted abundance distribution for mobile mRNA only, 95% confidence intervals shown, calculated using the Clopper-Pearson Binomial method. The predicted distributions fit the experimental data well. This suggests that the observed experimental results could predominantly be the result of passive filtering processes, without widespread sequence specific regulation of mobility. More low abundance transcripts are experimentally mobile than can be explained by the abundance model. This suggests that a secondary mechanism also affects mobility. It is not thought likely that the putative secondary mechanism only acts on low abundance transcripts, but instead that at other abundance values its effect is obscured by transcripts made mobile by the processes described in the abundance model.

the low abundance mobile transcripts ( ln(abundance) $< 1$, left hand side of Figure 4.5a). These transcripts are listed in Table 4.2, in Appendix. Whereas for the full dataset we failed to find any statistically significant motifs, for this subset we identified 3 statistically enriched motifs, (listed in Table 4.1), using DREME [8]. Analysis of Gene Ontology terms revealed an enrichment of processes associated with defence response and the chloroplast for this subset of transcripts (see Table 4.3, in Appendix).

Table 4.1: Putative mobility motifs identified in the low abundance mobile transcripts using DREME software [8]. The motifs were enriched in the mobile vs. non-mobile low abundance transcripts. See Table 4.2 in Appendix for the list of transcripts. Data from Thieme et al. [219]. P-value is calculated using Fisher's Exact Test, E-value is the P-value multiplied by the number of candidate motifs tested. Motifs with E-value $< 0.05$ are shown.

| Enriched motif | P-value | E-value |
|---|---|---|
| AGTWCAAC | 7.6E-7 | 2.8E-2 |
| ATGGTTTG | 8.7E-7 | 3.2E-2 |
| CCCACS | 1.3E-6 | 4.7E-2 |

### 4.2.4 Regulation of mobility through control of abundance proximal to the vasculature

It is possible that local transcript abundance near sieve elements is altered relative to the rest of the tissue to control movement from the site of transcription into sieve elements, and thus, to regulate mRNA mobility. To investigate this possibility, we analysed two available data sets, one with bundle sheath data [7] and the other with companion cell data [147]. In the bundle sheath, the mobile population was not enriched relative to overall leaf expression levels (Figure 4.6a, Figure 4.7). However, using the more localised companion cell data, we found that mobile mRNA transcripts were slightly but significantly over-expressed relative to the rest of leaf (Figure 4.6b, Figure 4.7). This suggests that local regulation of abundance may be a plausible mechanism for regulating mRNA mobility, although we do not see clear evidence that it definitely is.

### 4.2.5 mRNA half-life contributes to transcript mobility

The abundance model predicts that mRNA half-life should affect the probability of mobility, as more stable transcripts are likely to have more chances to move out of the cell into sieve elements before decaying (Figure 4.8. Consistent with this expectation, the mobile population had a greater half-life than the non-mobile population (Figure 4.9).

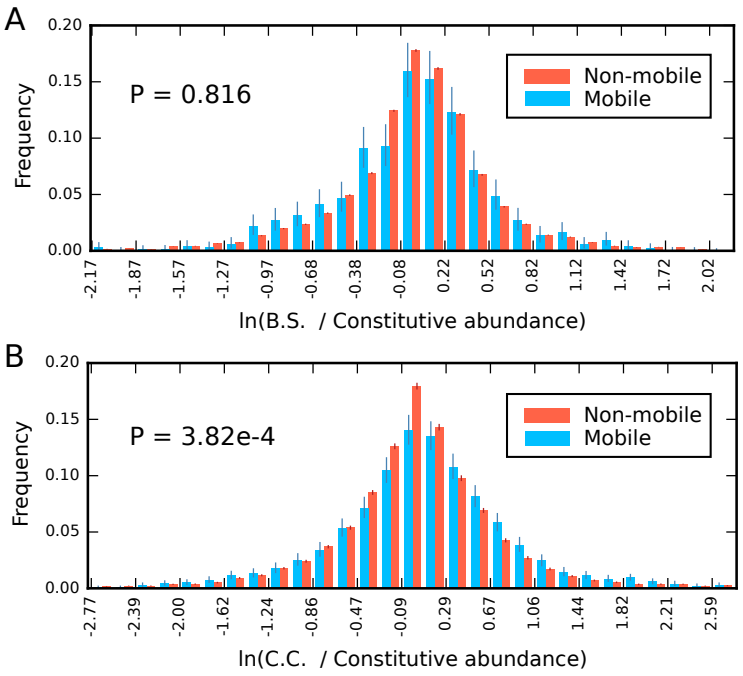



Figure 4.6: Mobile transcripts may be preferentially expressed proximal to sieve elements. The transcript expression ratio in cells proximal to the sieve element relative to the rest of the leaf: A, in the bundle sheath (B.S.), B, in companion cells (C.C.). The statistical significance of the difference of the means, P-value, was calculated using Welch's t-test. Abundance data taken from Mustroph et al. [147], Aubry et al. [7], and Thieme et al. [219], mobility classification from Thieme et al. [219]. Abundance ratio distributions in the bundle sheath, and companion cell are similar for mobile, and non-mobile transcripts, i.e. mobile transcripts are not generally overexpressed proximal to the vasculature relative to the rest of the leaf. We therefore do not see evidence to suggest widespread regulation of abundance proximal to the vasculature in order to regulate mobility, although it is certainly not eliminated as a potential regulatory mechanism.

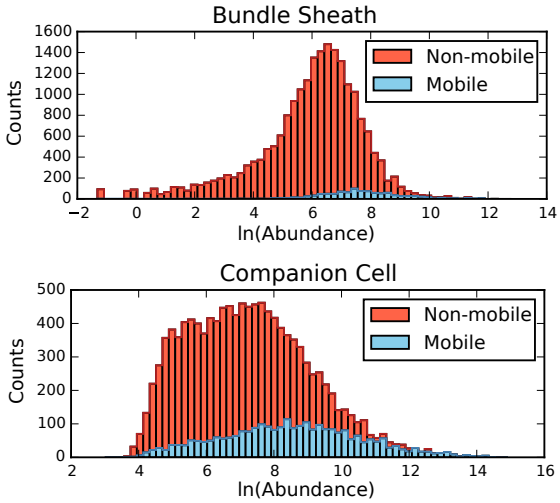



Figure 4.7: The abundance distribution of mobile and non-mobile transcripts in cells proximal to the vasculature. Abundance data for the bundle sheath from Aubry et al. [7], for the companion cell from Mustroph et al. [147], mobility classification data from Thieme et al. [219]. Transcripts which were found to be mobile in the Thieme dataset are relatively abundant in these datasets as well.

However, abundance is a function of transcription rate and half-life, and so this difference could be due to the effect of half-life on abundance, rather than the separable effect predicted by the model. To address this question, we performed linear discriminant analysis to find the most informative projection of the data to separate mobile from non-mobile transcripts, and found that the dominant eigenvector was

$$(\mathrm{v}_{\mathrm{abundance}}\ ,\ \mathrm{v}_{\mathrm{half\text{-}life}}) = (0.992, 0.123)$$

indicating that there was a half-life effect on mobility separable from its effect on abundance, but that this contribution was small relative to the size of the abundance effect. Visually, the best boundary to discriminate mobile from non-mobile transcripts found by logistic regression could be seen to have both an abundance and a separate half-life component (Figure 4.10).

### 4.2.6 Smaller transcripts appear to be more mobile

The observations discussed to this point could also be explained by detection sensitivity in the RNA-sequencing experiment; if more mRNA species are truly mobile than were experimentally detected, one might expect the more abundant, and more stable transcripts to be more likely to be detected, than others. We see that a toy detection threshold model (see Methods) results in a similar relationship between abundance and the probability that a transcript is mobile

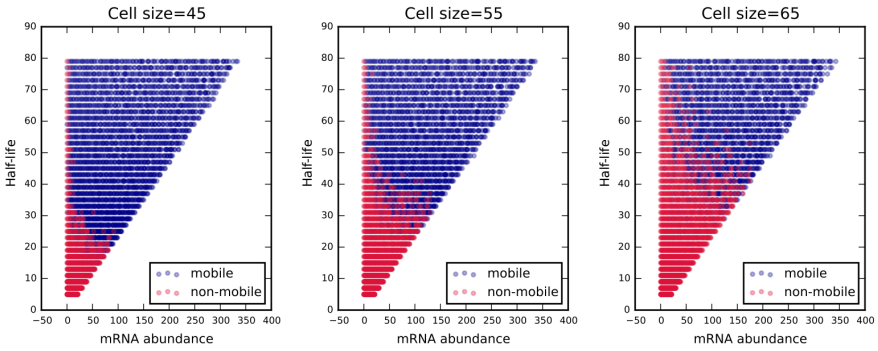



Figure 4.8: The predicted effect of half-life and abundance on transcript mobility. For varied transcription rates and half-lives, transcript abundance and mobility was simulated using the abundance cell model. Transcripts with longer half-lives can be seen to be more mobile than shorter transcripts. However, the effect of half-life on mobility can be seen to interact with transcript abundance, such that the threshold half-life, and sensitivity to changes in half-life varies with species abundance.

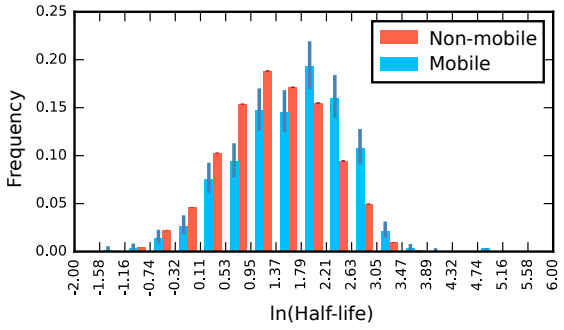



Figure 4.9: The distribution of half-life for experimentally determined mobile and non-mobile mRNA populations. Data taken from Narsai et al. [148] and Thieme et al. [219]. 95% Clopper-Pearson binomial confidence intervals shown. Mobile transcripts can be seen to be generally more stable than non-mobile transcripts, in agreement with the abundance model.

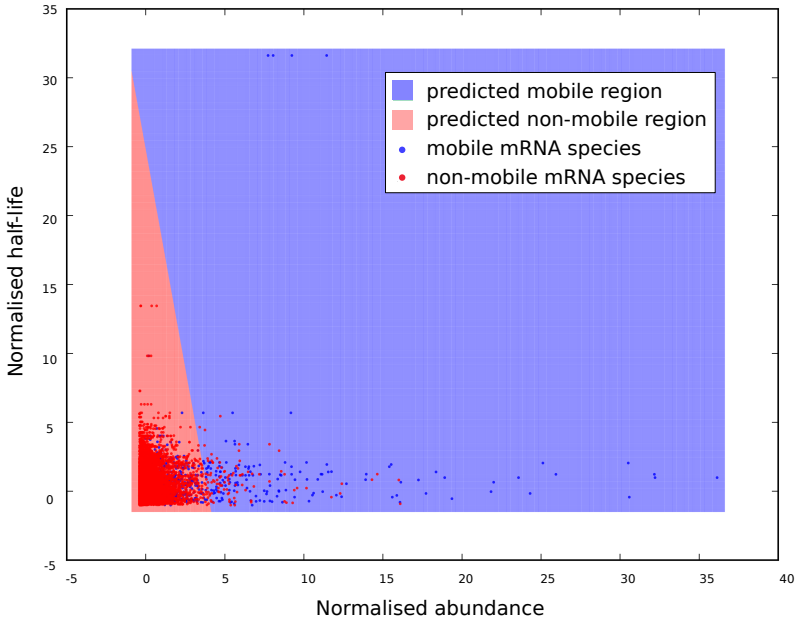



Figure 4.10: The contributions of half-life and abundance to mRNA mobility. The relationship between experimental half-life, abundance and mRNA species mobility is shown, and the predicted regions of mobility, and non-mobility, generated using a logistic regression classifier. Abundance and half-life were both normalised such that the mean value is 0 and standard deviation is 1. The boundary between the predicted regions has both abundance, and half-life components, suggesting that half-life has an effect on mobility, independent from its effect on abundance, as expected under the abundance model.

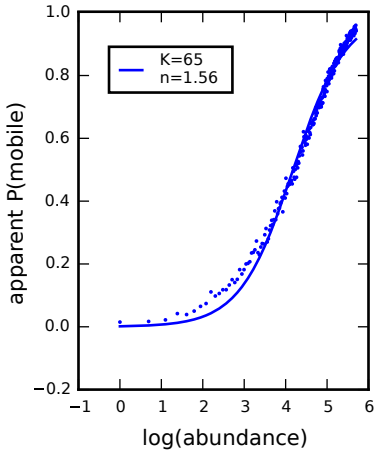



Figure 4.11: An experimental detection threshold model results in a similar abundance versus probability mobile curve as the abundance model. Points correspond to the estimated probability that a truly mobile mRNA of the indicated abundance is successfully detected as mobile using the detection threshold model (see Methods). The line fitted to the point data is of the same form as Equation 4.1. The good fit indicates that the abundance versus probability mobile curve cannot easily be used to distinguish between these models.

as the abundance model, which can be well described by the same saturation equation (Figure 4.11).

To distinguish between a detection threshold explanation, in which the effect of half-life, and abundance on mRNA mobility are an experimental artefact and the abundance model, in which abundance and half-life modulate the probability of escaping the producing cell, we considered the effect of molecule size on mobility.

Under the abundance model, transcripts with a larger Stokes radius would be less likely to be mobile, as they are slower to diffuse within a cell, and within a given time less likely to reach plasmodesmata. Although complicated by the formation of RNA secondary structures, we considered transcript length as a proxy for the Stokes radius of an RNA species. The dependence of transcript abundance in the non-producing distal tissue as a function of transcript length is shown in Figure 4.12a. The small but statistically significant negative correlation qualitatively supported that larger transcripts are less mobile. To check that this was not due to experimental detection bias, we analysed the dependence of local transcript abundance in the mRNA producing tissue as a function of transcript length. We would expect that experimental bias to be similar in local and distal tissue; however, we did not observe this (Figure 4.12b). By contrast to Figure 4.12a, we found no negative correlation between mRNA transcript length and local abundance (a minor positive correlation was observed), suggesting that experimental bias does not cause the size effect tendency.

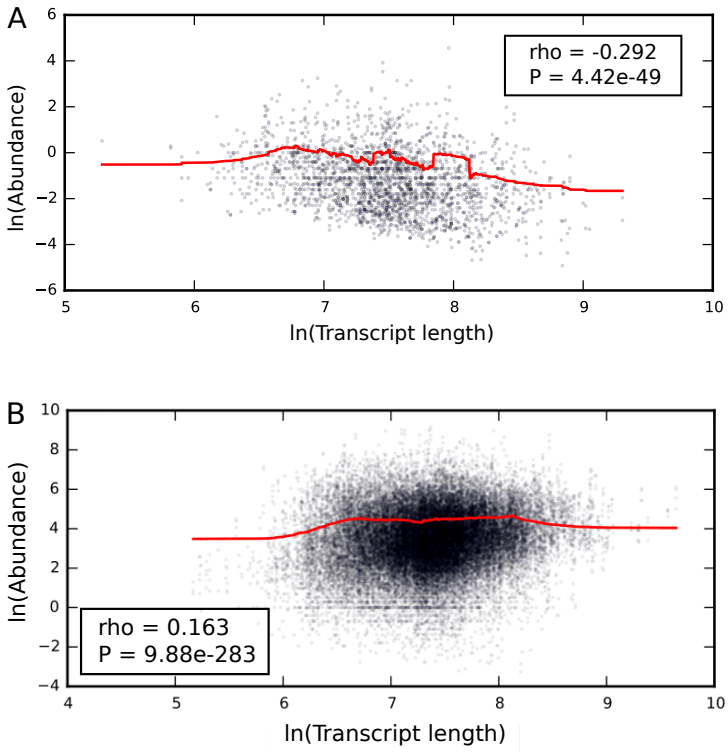



Figure 4.12: The effect of transcript length on mobility suggests that the observed trends in mobile mRNA species are at least partially caused by the abundance, rather than detection threshold model. A, Smaller transcripts are more mobile. This plot shows the mRNA abundance in the distal tissue, (i.e. only for mobile transcripts), as a function of transcript length. Data taken from Thieme et al. [219]. P-values were computed from Spearman's rank correlation, the moving average (red) was calculated with a window size of 300. The negative correlation indicates that smaller transcripts are more abundant in the distal, non-producing tissue, in agreement with the abundance model. B, Detected transcript abundance as a function of length in the producing tissues. A small but statistically significant positive correlation indicates that there is a slight detection bias favouring longer mRNA transcripts. This indicates that the negative correlation seen in Figure 4.12a is not a consequence of experimental sequencing bias leading to misrepresentation of transcript species abundance. Data taken from Thieme et al. [219], rho and p-values were calculated using Spearman's rank. The moving average (red) was calculated using a window size of 6,000.

## 4.3 Discussion

Using a simple computational model, we have shown that the large mobile mRNA population recently identified by Thieme et al. [219] can be explained by non-sequence- specific movement of mRNA into sieve elements. Within this model, mRNA abundance is a key determinant of mobility. Furthermore, we have shown that mRNA half-life and transcript length affect the mobile mRNA population in a manner consistent with the abundance model. Recently, the apparently non-specific loss of proteins from companion cells into the sieve-elements has been observed [159], suggesting that this could be a wide-spread phenomenon among diverse classes of molecules.

The consistency of the abundance model with existing experimental data does not imply that identified mobile mRNA species are not biologically relevant signalling molecules. The probability of reaching the cell surface itself could be a biologically relevant and regulated mechanism, in which the balance between half-life and transcription rate determines the mobility of mRNA species, indeed mRNA 5' and 3' UTR regions associated with increased mobility have been shown to increase transcript half-life [10]. Interestingly, the motif 'ATGGTTTG' which was enriched in low abundance, mobile transcripts (Table 4.1) has been previously found to be associated with stable transcripts [148], suggesting that stability can indeed compensate for low transcript abundance as predicted by the model in Figure 4.8. This is also supported by the discovered association of low abundance, mobile transcripts with the chloroplast, as shown in 'annotation cluster 2' of Table 4.3. Transcripts encoding chloroplast proteins have previously been shown to have a significantly higher proportion of transcripts with long half-lives [148].

Although we have predominantly used tissue-level expression data, mobile transcripts are also highly abundant in companion cells relative to non-mobile (Figure 4.7). Furthermore, mobile transcripts are slightly overexpressed in companion cells relative to constitutive expression, although it remains unclear whether this is evidence for a regulatory process governing mobility or whether mobility is a side effect of a transcriptome that has been changed for other purposes.

Our model defines mRNA mobility as the escape probability from companion cells and does not explicitly consider the transport process through sieve elements. This does not rule out a possible sequence-specific unloading process. Experimental data suggest that once a molecule is in the sieve elements it can move bi-directionally across a graft junction [126]. Therefore, we did not impose any directionality of mRNA movement within sieve elements flow. Should quantitative measures for transcript movement with sieve elements become available, the model could be readily extended to include this information.

Key to reproducing experimentally determined mRNA mobility from the model is the saturation curve shown in Figure 4.1b and Figure 4.5a. Our proposed abundance model, explained by transcription rate, diffusion and half-life, nat-

urally captures this behaviour. However, we point out that, in principle, any process that gives rise to such an abundance-mobility saturation curve could explain the data.

In developing the presented abundance model, we made a number of approximations that likely warrant future extensions, such as not including advection in cytoplasmic transport and using simple box-shaped cells. Notably, we have not needed to account for different transport probabilities through the plasmodesmata, say, as a function of mRNA size or shape to explain the data. Analysis of the experimental data shows a correlation of abundance in the distal tissue with transcript length, but does not reveal a size threshold, which would be indicative of a size exclusion limit. Given the size of mRNA transcripts, they could be actively chaperoned through the plasmodesmata, perhaps similarly to viral RNA.

If mRNA transport through plasmodesmata requires chaperones that recognise a sequence motif that binds with an equilibrium dissociation constant, $K_d$, then mRNA with a concentration close to or above that $K_d$ would bind and be transported. A transcript that has a different but similar motif may result in weaker binding that would require a higher abundance to bind. Thus, selective and non-selective mRNA mobility may be conceivably part of a common transport process, with abundance as the determining factor. The presented model does not exclude the possibility of mRNA motifs playing a role in mobility. The tight relationship, however, between mRNA mobility and abundance for the bulk of the available data clearly demonstrates the importance of abundance, whereas a similar relationship between mobility and sequence motifs across a large range of transcripts has yet to be shown. The few putative motifs we identified in a reduced data set require further experimental investigation.

An alternative mechanism for mRNA movement could be one in which mRNA enters sieve elements from sieve tube precursor cells, which undergo partial apoptosis during differentiation. If this were the case, all mRNA could move across the graft junction and it is possible that experimental detection bias of mRNA might potentially give rise to the above-mentioned abundance-mobility saturation curve. However, this possibility is not supported by the trends in the transcript length and count data, which were not consistent across tissues. Furthermore, the implications of this mechanism are the same as for the presented abundance model in that the detected mobile mRNA transcripts are mobile predominantly because of abundance rather than sequence-specific transport processes.

Based on our results, we suggest that the large majority of identified transcripts are unlikely to be selectively transported. However, it is possible that mRNA species made mobile through the processes described in the abundance model obscure a relatively small population that is made mobile through a different mechanism, as evidenced by mRNA fusion studies performed by Thieme et al. [219] and the two statistically enriched motifs identified in the low-abundance mobile population which have not been linked to stability. We propose that

the presented abundance model should be considered the null hypothesis when assessing mRNA mobility data and other mechanisms of mRNA transcript mobility.

Since this work was carried out, a study by Zhang et al. [254] identified tRNA derived motifs associated with mRNA mobility. They argue for an "active and regulated" mRNA delivery mechanism on the basis of 1.) the presence of these sequences, 2.) the transfer of mRNA to "specific aboveground tissues" , and 3.) that mobile mRNA does not necessarily follow the source to sink flow of phloem contents.

However, it has been shown previously that molecules are able to move bidirectionally in the phloem, and cross graft junctions both with, and against the net flow in the phloem [126]. Consistent with bidirectional movement of mRNA within the phloem, Zhang et al. observed reduced mobility against the phloem source to sink phloem flow direction [254], suggesting that there may be no need to invoke the proposed "active transport system".

Secondly, although mRNA species could be interpreted as exhibiting specific movement into particular tissues, this observation is also consistent with the stochastic movement implied by the abundance model. When transcripts cannot be consistently identified as mobile across biological replicates even in tissues as comparatively large as the rosette and the root system [219, 254], it is not surprising that in a single experiment a given transcript species is "specifically" identified in some aboveground tissues and not others.

The identification of these mobility associated motifs is an exciting step, however, they are not predicted to interact with polypyrimidine tract binding proteins found in the phloem [254], and their mode of action remains unclear. The sequences are often found in the 3'-UTR, which is well known to be associated with mRNA stability (reviewed [143]), and it is possible that their effect is mediated through mRNA stability. This may explain the somewhat stochastic nature of their mobility in engineered transcripts (n=9/44, n=6/25 for the two discovered motifs [254]). Unfortunately, no stability assay was carried out to evaluate this effect.

The identified motifs were associated with only 11.4% of the identified mobile transcripts [254], similar in magnitude to the number of unexplained transcripts under the abundance model Figure 4.5c, and it remains an interesting, and I think largely unresolved question, as to what extent mRNA mobility plays an important role in plant physiology, and the mechanisms by which it is achieved.

## 4.4 Methods

### 4.4.1 Data sources

Abundance and mobility data was taken from Supplementary Information 1 of Thieme et al. [219]. Thieme et al. [219] performed a grafting experiment, in which two distantly related ecotypes of *Arabidopsis thaliana*, Ped-0 & Col-0, dispaying a high frequency of genomic sequence single nucleotide polymorphisms (SNPs) were grafted together. RNA-sequencing of each half of the grafted plants was performed, and ecotype specific SNPs present in RNA molecules allowed them to identify their origin, and thus their mobility in reciprocal chimaeric root-shoot grafted plants.

Transcripts with less than three read counts were excluded from the data. For each transcript, in each grafted tissue, 'abundance' was calculated as the average read count per informative SNP in the local, producing tissue. Transcripts were considered 'mobile' if the read count for the non-local form of the transcript was greater than zero in the reciprocal grafted tissue. mRNA half-life data was taken from Supplementary Table 2 of Narsai et al. [148]. Companion cell, and bundle sheath abundance information used in Figure 4.6, and Figure 4.10 was taken from supplementary data of Mustroph et al. [147] and Aubry et al. [7].

### 4.4.2 Calculation of escape probability from many cells

Probability of escape from a single cell, $P(E)$, was computed using the abundance model as described in subsection 4.2.1. The expected probability of a mRNA molecule moving into sieve elements from multiple companion cells, $P(F)$ was be calculated as $P(F) = 1 - (1 - P(E))^m$ , where $m$ is the number of companion cells.

### 4.4.3 Mobility prediction & fitting to abundance data using saturation curve

The saturation curve equation used to describe the predicted relationship between mRNA mobility and abundance was

$$P(m) = \frac{A^n}{K^n + A^n} \tag{4.1}$$

where $P(m)$ is the probability of the transcript being mobile, $A$ is the experimentally measured transcript abundance, $K$ is the abundance for which the

probability of being mobile is 50%, and $n$ gives the steepness of the curve. For each modelled transcript $t$, of abundance $A$, in the set of all experimentally measured transcripts ($T$), $P(m)$ was calculated, using Equation 4.1, and $A_t$ was exclusively assigned to either the mobile set $M$, or the stationary set $S$ based on the value of $P(m)$ such that

$$\begin{aligned} P(A_t \in M) &= P(m) \quad \forall\, t \in T \\ M \cap S &= \emptyset \\ M \cup S &= \{A_1, ..., A_T\}. \end{aligned}$$

This reflects the experimental approach taken by Thieme et al. [219] in which transcripts where classified as either 'mobile', or 'non-mobile' depending on whether they were detected in the non-producing half of the grafted plant.

To fit the $K$, and $n$ parameters to the experimental data, the distribution of abundances for predicted and experimental mobile and non-mobile transcripts were approximated by histograms, and the difference between the area of predicted and experimental distributions minimised using the `scipy.basinhopping` algorithm.

The Clopper-Pearson method was used to calculate the binomial confidence interval for the proportion of the population expected to be mobile in each abundance range, as this allows the calculation of confidence intervals even for abundance ranges in which no mobile transcripts were detected / predicted.

### 4.4.4 Detection threshold model

Under the assumption that all mRNA molecules are mobile, RNA-sequencing sensitivity could produce a similar relationship between abundance and detected probability of mobility. To model the effect of detection sensitivity we sampled from a binary array $R$. The number of elements in $R$ corresponds to the number of all mobile transcript molecules present in the non-local tissue, the number of non-zero elements correspond to the number of mobile elements of a particular transcript species of interest. Under the assumption that local cell and tissue geometry provides no differential impediment to mobility, this is linearly related to mRNA abundance in the producing tissue.

For each mRNA abundance, the probability of successfully detecting that a transcript was mobile, apparent P(mobile), was estimated by sampling from $R$, 10,000 times, and calculating the fraction of samples containing a non-zero element. Sample size is equivalent to the fraction of all mobile mRNA molecules which were sequenced.

### 4.4.5 Linear discriminant analysis

Linear discriminant analysis is a method, similar to principle component analysis, to find a linear combination of features which best separates two predefined classes, here mobile, and non-mobile transcripts. Here we used the implementation in `scikit-learn` (v0.17.1) to demonstrate that the principle dimension includes both half-life, and transcript abundance elements, and that therefore the effect of half-life on mobility was not only caused by the effect of half-life on abundance.

### 4.4.6 Logistic regression

Logistic regression is a commonly used tool in applied statistics, in which the logistic transformation of the probability of set membership is modelled by a linear equation

$$log\frac{P(x)}{1-P(x)} = \beta_0 + x_1 \cdot \beta_1 + x_2 \cdot \beta_2. \tag{4.2}$$

The implementation of logistic regression in `scikit-learn` (v0.17.1) was used to examine the relationship between abundance, half-life, and mobility, such that $P(x)$ was the probability that a transcript is mobile, $x_1$ was normalised half-life, and $x_2$ was normalised abundance.

## 4.5 Appendix

Table 4.2: List of low abundance mobile transcripts present in the data set of Thieme et al. [219]. To define low abundance we took a threshold based on Figure 4.5a, $\ln(\text{abundance}) < 1$.

| transcript ID | Gene Symbol | Gene Model Description |
|---|---|---|
| AT1G16160.1 | WALL ASSOCIATED KINASE-LIKE 5 (WAKL5) | WAK-like kinase |
| AT1G21250.1 | CELL WALL-ASSOCIATED KINASE (WAK1) | cell wall-associated kinase, may function as a signaling receptor of extracellular matrix component such as oligogalacturonides. |
| AT1G32060.1 | PHOSPHORIBULOKINASE (PRK) | phosphoribulokinase (PRK) |
| AT1G42970.1 | GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE B SUBUNIT (GAPB) | chloroplast localized glyceraldehyde-3-phosphate dehydrogenase |
| AT1G52000.1 | | Mannose-binding lectin superfamily protein |
| AT1G52400.1 | BETA GLUCOSIDASE 18 (BGLU18) | member of glycosyl hydrolase family 1, located in inducible ER bodies, required in inducible ER body formation |
| AT1G56510.1 | WHITE RUST RESISTANCE 4 (WRR4) | TIR-NB-LRR protein that confers resistance to four races of Albugo candida. |
| AT1G61520.1 | PHOTOSYSTEM I LIGHT HARVESTING COMPLEX GENE 3 (LHCA3) | PSI type III chlorophyll a/b-binding protein (Lhca3*1) |
| AT1G61610.1 | | S-locus lectin protein kinase family protein |
| AT1G65800.1 | RECEPTOR KINASE 2 (RK2) | encodes a putative receptor-like serine/threonine protein kinases that is similar to brassica self-incompatibility (S) locus |
| AT1G69730.1 | | Wall-associated kinase family protein |
| AT1G75920.1 | | GDSL-like Lipase/Acylhydrolase superfamily protein |
| AT2G35370.1 | GLYCINE DECARBOXYLASE COMPLEX H (GDCH) | Encodes glycine decarboxylase complex H protein |
| AT2G44840.1 | ETHYLENE-RESPONSIVE ELEMENT BINDING FACTOR 13 (ERF13) | encodes a member of the ERF (ethylene response factor) subfamily B-3 of ERF/AP2 transcription factor family |
| AT3G02620.1 | | Plant stearoyl-acyl-carrier-protein desaturase family protein |
| AT3G08940.2 | LIGHT HARVESTING COMPLEX PHOTOSYSTEM II (LHCB4.2) | Lhcb4.2 protein involved in the light harvesting complex of photosystem II |
| AT3G22121.1 | | Potential natural antisense gene, locus overlaps with AT3G22120 |
| AT3G26610.1 | | Pectin lyase-like superfamily protein |

Table 4.2: List of low abundance mobile transcripts present in the data set of Thieme et al. [219]. To define low abundance we took a threshold based on Figure 4.5a, $\ln(\text{abundance}) < 1$.

| transcript ID | Gene Symbol | Gene Model Description |
|---|---|---|
| AT3G44860.1 | FARNESOIC ACID CARBOXYL-O-METHYLTRANSFERASE (FAMT) | Encodes a farnesoic acid carboxyl-O-methyltransferase. |
| AT3G46780.1 | PLASTID TRANSCRIPTIONALLY ACTIVE 16 (PTAC16) | plastid transcriptionally active 16 (PTAC16) |
| AT3G49110.1 | PEROXIDASE CA (PRXCA) | Class III peroxidase Perx33 |
| AT3G50300.1 | | HXXXD-type acyl-transferase family protein |
| AT3G55800.1 | SEDOHEPTULOSE-BISPHOSPHATASE (SBPASE) | Encodes the chloroplast enzyme sedoheptulose-1,7-bisphosphatase (SBPase), involved in the carbon reduction of the Calvin cycle |
| AT3G61270.1 | | best Arabidopsis thaliana protein match is: downstream target of AGL15 2 (TAIR:AT2G45830.1) |
| AT4G11000.1 | | Ankyrin repeat family protein |
| AT4G13130.1 | | Cysteine/Histidine-rich C1 domain family protein |
| AT4G13620.1 | | encodes a member of the DREB subfamily A-6 of ERF/AP2 transcription factor family |
| AT4G14630.1 | GERMIN-LIKE PROTEIN 9 (GLP9) | germin-like protein with N-terminal signal sequence that may target it to the vacuole, plasma membrane and/or outside the cell. |
| AT4G15160.1 | | Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein |
| AT4G19170.1 | NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 4 (NCED4) | chloroplast-targeted member of a family of enzymes similar to nine-cis-epoxycarotenoid dioxygenase |
| AT4G23150.1 | CYSTEINE-RICH RLK (RECEPTOR-LIKE PROTEIN KINASE) 7 (CRK7) | Encodes a cysteine-rich receptor-like protein kinase. |
| AT4G26050.1 | PLANT INTRACELLULAR RAS GROUP-RELATED LRR 8 (PIRL8) | Encodes PIRL8, a member of the Plant Intracellular Ras-group-related LRRs (Leucine rich repeat proteins) |
| AT4G27440.1 | PROTOCHLOROPHYLLIDE OXIDOREDUCTASE B (PORB) | light-dependent NADPH:protochlorophyllide oxidoreductase B |
| AT4G29690.1 | | Alkaline-phosphatase-like family protein |
| AT4G31100.1 | | wall-associated kinase, putative |
| AT4G31354.1 | | This gene encodes a small protein and has either evidence of transcription or purifying selection. |

Table 4.2: List of low abundance mobile transcripts present in the data set of Thieme et al. [219]. To define low abundance we took a threshold based on Figure 4.5a, $\ln(\text{abundance}) < 1$.

| transcript ID | Gene Symbol | Gene Model Description |
|---|---|---|
| AT4G33610.1 | | glycine-rich protein |
| AT4G37220.1 | | Cold acclimation protein WCOR413 family |
| AT5G14650.1 | | Pectin lyase-like superfamily protein |
| AT5G25130.1 | CYTOCHROME P450, FAMILY 71, SUBFAMILY B, POLYPEPTIDE 12 (CYP71B12) | putative cytochrome P450 |
| AT5G25980.2 | GLUCOSIDE GLUCOHYDROLASE 2 (TGG2) | Myrosinase (thioglucoside glucohydrolase) gene involved in glucosinoloate metabolism. |
| AT5G26260.1 | | TRAF-like family protein |
| AT5G26280.1 | | TRAF-like family protein |
| AT5G28500.1 | | unknown protein |
| AT5G38980.1 | | unknown protein |
| AT5G44580.1 | | unknown protein |
| AT5G46330.1 | FLAGELLIN-SENSITIVE 2 (FLS2) | leucine-rich repeat serine/threonine protein kinase that is expressed ubiquitously |
| AT5G55930.1 | OLIGOPEPTIDE TRANSPORTER 1 (OPT1) | oligopeptide transporter |
| AT5G60760.1 | | P-loop containing nucleoside triphosphate hydrolases superfamily protein |

Table 4.3: Gene ontology terms associated with defence, and the plastid are enriched in the low abundance mobile transcripts relative to the low abundance non-mobile set. Enrichment analysis was carried out using DAVID 6.7 [92], which also clusters ontology terms together into more interpretable related annotation clusters.

| **Annotation Cluster 1** | **Enrichment Score 2.13** |
|---|---|
| *Term* | *P-value* |
| GO 0042742 defense response to bacterium | 2.64E-03 |
| GO 0009617 response to bacterium | 8.20E-03 |
| GO 0006952 defense response | 1.88E-02 |
| | |
| **Annotation Cluster 2** | **Enrichment Score 2.11** |
| *Term* | *P-value* |
| GO 0015979 photosynthesis | 2.73E-04 |
| GO 0009941 chloroplast envelope | 8.11E-04 |
| chloroplast | 9.15E-04 |
| GO 0009526 plastid envelope | 1.18E-03 |
| GO 0009579 thylakoid | 1.30E-03 |
| GO 0044435 plastid part | 1.57E-03 |
| GO 0044434 chloroplast part | 1.57E-03 |
| GO 0031975 envelope | 1.97E-03 |
| GO 0031967 organelle envelope | 1.97E-03 |
| transit peptide | 2.14E-03 |
| plastid | 4.12E-03 |
| calvin cycle | 4.80E-03 |
| GO 0009532 plastid stroma | 6.37E-03 |
| GO 0015977 carbon utilization by fixation of carbon dioxide | 6.52E-03 |
| GO 0019685 photosynthesis dark reaction | 6.52E-03 |
| GO 0019253 reductive pentose-phosphate cycle | 6.52E-03 |
| GO 0009507 chloroplast | 6.95E-03 |
| GO 0009536 plastid | 6.95E-03 |
| GO 0009535 chloroplast thylakoid membrane | 1.10E-02 |
| GO 0042651 thylakoid membrane | 1.10E-02 |
| GO 0034357 photosynthetic membrane | 1.10E-02 |
| GO 0055035 plastid thylakoid membrane | 1.10E-02 |
| transit peptide Chloroplast | 1.42E-02 |
| GO 0009534 chloroplast thylakoid | 1.96E-02 |
| GO 0031976 plastid thylakoid | 1.96E-02 |
| GO 0031984 organelle subcompartment | 1.96E-02 |
| GO 0044436 thylakoid part | 1.96E-02 |
| photosynthesis | 2.08E-02 |
| GO 0009570 chloroplast stroma | 3.57E-02 |
| ath00710 Carbon fixation in photosynthetic organisms | 3.59E-02 |
| GO 0048046 apoplast | 3.62E-02 |

| | |
|---|---|
| GO 0031090 organelle membrane | 6.83E-02 |
| GO 0006091 generation of precursor metabolites and energy | 7.00E-02 |
| GO 0005576 extracellular region | 8.00E-02 |
| GO 0016051 carbohydrate biosynthetic process | 1.35E-01 |

| **Annotation Cluster 3** | **Enrichment Score 1.99** |
|---|---|
| *Term* | *P-value* |
| disulfide bond | 1.84E-06 |
| disulfide bond | 1.30E-05 |
| IPR018097 EGF-like calcium-binding conserved site | 1.11E-04 |
| glycoprotein | 1.50E-04 |
| serine/threonine-protein kinase | 2.92E-04 |
| PIRSF000575 wall-associated protein kinase | 3.03E-04 |
| signal | 4.35E-04 |
| IPR017441 Protein kinase / ATP binding site | 7.14E-04 |
| glycosylation site N-linked (GlcNAc...) | 8.25E-04 |
| IPR008271 Serine/threonine protein kinase active site | 1.75E-03 |
| IPR017442 Serine/threonine protein kinase-related | 1.93E-03 |
| active site Proton acceptor | 2.22E-03 |
| IPR000719 Protein kinase core | 3.10E-03 |
| ATP-binding | 4.43E-03 |
| signal peptide | 4.59E-03 |
| domain Protein kinase | 4.66E-03 |
| region of interest Atypical EGF-like | 5.16E-03 |
| binding site ATP | 5.47E-03 |
| IPR013695 Wall-associated kinase | 5.49E-03 |
| kinase | 5.66E-03 |
| SM00181 EGF | 6.67E-03 |
| GO 0004674 protein serine/threonine kinase activity | 7.05E-03 |
| nucleotide-binding | 7.36E-03 |
| GO 0006468 protein amino acid phosphorylation | 7.47E-03 |
| GO 0005773 vacuole | 7.73E-03 |
| GO 0016310 phosphorylation | 8.08E-03 |
| GO 0006793 phosphorus metabolic process | 8.73E-03 |
| GO 0006796 phosphate metabolic process | 8.73E-03 |
| IPR006210 EGF-like | 8.97E-03 |
| GO 0004672 protein kinase activity | 1.32E-02 |
| nucleotide phosphate-binding region ATP | 2.05E-02 |
| membrane | 2.43E-02 |
| topological domain Extracellular | 2.69E-02 |
| GO 0005509 calcium ion binding | 2.76E-02 |
| transferase | 3.65E-02 |
| phosphotransferase | 4.55E-02 |
| transmembrane | 6.77E-02 |
| topological domain Cytoplasmic | 7.05E-02 |

| | |
|---|---|
| GO 0005524 ATP binding | 8.45E-02 |
| GO 0032559 adenyl ribonucleotide binding | 9.34E-02 |
| GO 0001882 nucleoside binding | 1.31E-01 |
| GO 0001883 purine nucleoside binding | 1.31E-01 |
| GO 0030554 adenyl nucleotide binding | 1.31E-01 |
| GO 0043169 cation binding | 1.38E-01 |
| GO 0043167 ion binding | 1.38E-01 |
| GO 0032553 ribonucleotide binding | 1.39E-01 |
| GO 0032555 purine ribonucleotide binding | 1.39E-01 |
| GO 0000166 nucleotide binding | 1.53E-01 |
| GO 0017076 purine nucleotide binding | 1.85E-01 |
| transmembrane region | 1.91E-01 |
| GO 0016021 integral to membrane | 3.43E-01 |
| GO 0031224 intrinsic to membrane | 4.76E-01 |
| GO 0005886 plasma membrane | 7.05E-01 |
| | |
| **Annotation Cluster 4** | **Enrichment Score 1.33** |
| *Term* | *P-value* |
| GO 0006952 defense response | 1.88E-02 |
| GO 0050832 defense response to fungus | 6.05E-02 |
| GO 0009620 response to fungus | 9.03E-02 |
| | |
| **Annotation Cluster 5** | **Enrichment Score 1.26** |
| *Term* | *P-value* |
| hydrolase | 3.04E-02 |
| region of interest Substrate binding | 4.22E-02 |
| binding site Substrate | 1.32E-01 |
| | |
| **Annotation Cluster 6** | **Enrichment Score 0.722** |
| *Term* | *P-value* |
| oxidoreductase | 1.22E-01 |
| GO 0055114 oxidation reduction | 1.30E-01 |
| iron | 4.32E-01 |
| | |
| **Annotation Cluster 7** | **Enrichment Score 0.652** |
| *Term* | *P-value* |
| GO 0009628 response to abiotic stimulus | 1.70E-01 |
| GO 0009416 response to light stimulus | 2.49E-01 |
| GO 0009314 response to radiation | 2.62E-01 |

# Chapter 5

# Discussion

Having recently reviewed work from all stages of the project, it now feels opportune to look back, and reflect on what has been a journey of personal and professional development, and, happily, some scientific discovery. To conclude this thesis, we discuss how, through the application of mathematical modelling, and statistical analysis, we have contributed to an understanding of the 'purpose' of metabolic flux patterns, and the extent to which this is an appropriate mindset to address biological questions.

## 5.1 Curation of metabolic model

Initially, we intended to build an ODE model of the kinetics of sulfur uptake and assimilation, building upon the specific expertise of the Kopriva group, and the Bayesian strengths of the Morris group in parameter estimation. Unfortunately, after several months, we demonstrated that insufficient experimental data was available to generate any meaningful result.

Following some discussion, it was established that carbon skeleton availability, rather than sulfur itself is often limiting to sulfur uptake under many environments. Consequently, the idea of largely parameter free, 'genome scale' models was appealing both scientifically, and practically, and therefore pursued. Although initially beyond the expertise of any project member, this relatively young field seemed full of opportunity, and early enthusiastic efforts lead to the results presented in chapter 2, as well as providing the chance to study MATLAB.

As newcomers to the field of constraint-based approaches, we were initially taken back by the relatively poor agreement of many published models of plant metabolism, with basic experimental data. Although these limitations were not

necessarily obvious based upon the analysis carried out in the original publication, further investigation often yielded surprising errors. This is due not so much to difficulty of the approach, as to the sheer scale of the modelling problem. Although simplifying assumptions regarding steady state are made, the enormous complexity of plant metabolism, and the resulting huge number of reactions which must be accounted for, make the production of these models a hugely time consuming task, especially in relation to smaller ODE based approaches. As described in chapter 2, we have spent a considerable amount of effort in further refining one such model. This resulted not only in a greater agreement with the qualitative predictions of gene requirements, which we were explicitly fitting the model to, but also to a slight improvement in flux predictions as compared to tracer experiments. This model has also provided the basis of much of the rest of the work carried out.

It is interesting that in similar models of human metabolism, considerable collaborative effort has gone into attempting to maintain a single, consensus model, and recently into re-integrating parallel models [214]. Conversely, in Arabidopsis, and plant metabolism generally, a number of separate models have been maintained, in parallel, across a considerable number of years, and publications. Given how time consuming these models are to make, the maintenance of parallel models clearly results in a duplication of effort, and is therefore undesirable.

It is unclear what, if any, underlying differences in understanding, or approach have led to these differences in outcome between the study of human and Arabidopsis metabolism. We speculate that the increased complexity of plant in comparison to human metabolism may lead to increased uncertainty, and an unwillingness to accept a consensus model. However, we observed that no two of the Arabidopsis models evaluated, use the same nomenclature for identifying model components. Consequently, in spite of the common use of SBML, it is difficult to compare the specific differences between models, to evaluate metabolic regions of similarity and difference, in order to determine whether they are controversial. Although the integration of the various published models of Arabidopsis is an important step in the maturation of the field, it is a complex task, and expected to require extensive work. However, the enforcement of common (data) standards is increasingly acknowledged to be an important step across various aspects of biology, as such it is strongly to be desired that a common system of agent identification be adopted.

It is desirable to evaluate genome-scale metabolic modes by a common metric. In work published to date, model quality is assessed primarily in relation to the interests of the authors. Whilst this is understandable, it limits the reuse of previously published models. We propose that the ability of a model to produce biomass from the commonly available inorganic metabolites, and comparison of gene lethality predictions to the Lloyd & Meinke database [128] be reported. Although not sufficient to completely describe the vagaries of a given model, this at least allows some standardised comparison of the quality of the

published Arabidopsis models. Previously these have not necessarily been appropriate as metrics, as particular models have focused on relatively small areas of metabolism, and therefore permitted the external supply of organic metabolites. However, given that models have now been published which are able to produce biomass from inorganic substrates, there is no reason that further, more focused models should not be integrated into these.

In spite of the availability of a number of high quality $^{13}$C-MFA flux measurement datasets in Arabidopsis, we do not suggest the use of flux predictions for general use in comparison of relative model quality. Measurements are limited to only small metabolic regions relative to the whole of metabolism, therefore although they should of course be considered for verification of the particular metabolic regions of interest, they do not generally have broad coverage across the model. Of course transcriptomic datasets offer better coverage, but are their use to infer flux is controversial. Furthermore, flux predictions derived from the model are somewhat sensitive to the particular method of constraint-based analysis used, and thus may reflect this, rather than the quality of the model itself.

## 5.2 FBA summary

Flux balance analysis is a computationally tractable method for predicting flux distribution in a genome scale model. We have used FBA to attempt to address the high error rate of current unbiased gene identification approaches, by providing a second, independent means to identify important reactions for the production of glucosinolates, important secondary metabolites. The idea being that two approaches could then be used in conjunction to preferentially target genes identified by both methods.

We found that indeed the model was able to predict genes which are involved in glucosinolate production with greater accuracy than other unbiased approaches. This serves to partially validate the model, and in particular, the areas of secondary metabolism which have not previously been assessed. However further investigation suggested that the FBA approach cannot be considered as unbiased as initially assumed; although it nominally considers all of metabolism, in fact the quality of predictions is not equal across all genes which in fact affect glucosinolates, and is biased towards the successful recovery of already expected genes. Consequently, it is not clear that FBA can currently be usefully applied to facilitate the understanding of secondary metabolism in this way. However, we did see that interestingly, FBA appears to be able to make predictions of enzymes which have more subtle effects than is commonly achieved in GWAS studies. Consequently, we speculate that as the quality of published models continues to improve, FBA based methods will one day be usefully applied to the engineering of plants, as has already been done in microbes.

A quirk of the FBA method, is that it only returns optimal solutions. We have previously discussed this as a limitation for flux prediction, however it can be interesting to observe how evolution has guided a plants behaviour to an optimality, and therefore to help understand the 'motivation' behind the way that metabolism responds to perturbation. In a simple model, integrating kinetic, and constraint-based approaches, we were able to see that an essentially biphasic response to slight and severe sulfur stress is the optimal solution to maintain the ability to produce biomass, and that this is a consequence of the non-linear relationship of transport to substrate concentration. When we extended this study to explicitly consider the changes in internal fluxes, we saw good agreement between model predictions of the response of key enzymes to sulfur starvation and experimentally observed changes, however, overall the level of agreement was somewhat middling when compared to a full transcriptomic dataset.

The difference between discussed, and all studied genes is striking. It is interesting how, often, little of the data generated in 'omic studies can be worked into an interesting, biological narrative, and discussed within the body of the paper. Even when an initial hypothesis is answered, the huge datasets generally contain a great deal more information. However, it is often extremely difficult to work backwards from the data, either to infer regulatory networks, or to understand to what 'purpose' metabolism is being regulated. This must be ascribed to the complexity of the networks of metabolic reactions, and the regulation of these reactions.

Our results again tend to suggest that FBA is currently well able to explain the expected, and easily discussed genes rather than all truly relevant genes. However, the potential, (as models continue to improve), of constraint-based methods to understand the motivation behind changes in gene expression levels in plant in response to environmental stresses, in order to generate an explanatory story is clear.

Transcriptomic datasets are hugely abundant, and offer by far the best coverage, not only of an enormous variety of perturbations, but also of the reactions of the metabolic network. However, it should also be remembered that transcript abundance is not an ideal experimental indicator of flux through an associated reaction; correlation between transcript, and protein abundance is generally fairly weak, and it is expected that correlation to flux will be even weaker. It is currently difficult to pinpoint which disagreements are due to model error, and which are due to the use of somewhat inappropriate data.

Although I think we have seen that not every conclusion of FBA is supported by the data which are available, and that often the ideal dataset with which to test these ideas does not exist, the output of the FBA based modelling approach is often consistent with the data, and allows the generation of ideas, in particular with regards to the 'purpose' behind metabolism, which I do not think would easily be possible through other means. Plant metabolism is, of course, a hugely complex subject. Although not a new approach, constraint based modelling was

a new idea to me, it is fantastically interesting, and I think surprising that even some aspects of such a complex system can be addressed through the application of such a simple framework.

## 5.3 EFMs summary

At around the time this work was performed, Thomas Wilhelm was working on a method for the calculation of diverse EFM sets, with Joern Behre at the Institute of Food Research, in Norwich. This was extremely exciting, and potentially provided a powerful method for the analysis of our recently updated model using a reduced set of elementary modes. Furthermore, Thomas and Joern are both experienced in constraint-based modelling approaches, and their collaboration allowed us to address the lack of expertise within our group. Unfortunately, although a large codebase implementing their method had been written by a succession of developers, it could not be practically applied, due to the computational resources required. After extensive work, learning enough Java to improve the code, analysing bottlenecks, and tweaking the parameters of the Cplex Optimiser software, it still could not be practically used, highlighting the computational difficulties of an EFM based approach. However, the publication of the more efficient TreeEFM method [163] has allowed the application of EFMs to the model, and, as presented in chapter 3, generated a number of hypotheses concerning various aspects of plant nutrition. It is interesting to note that although modelling is normally cited as a means to target, and thus reduce experimental requirements, within the sphere of EFMs, computational requirements can potentially be alleviated through additional experimental work; metabolite concentration data can be used to reduce the requirement for EFM calculation by eliminating thermodynamically inconsistent modes.

Elementary flux modes provide an efficient description of the capabilities of a metabolic network. Although it remains impractical to calculate all elementary modes, we have seen that a subset of the modes can be calculated for genome scale models, and that this subset can be used to approximate the behaviour of the full set. We have applied EFMs to study nutrient use efficiency. Interestingly this has highlighted the importance of considering waste metabolites carefully. The importance of these exchange fluxes has perhaps been neglected in many previously published models, in favour of the biomass equation, and uptake fluxes.

Yield space analysis of the elementary modes has generated a large number of ideas about nutrient use efficiency. Personally I had expected to see a greater potential for trade-offs between the requirement for different nutrients, which appears relatively insensitive to 'decisions' available to the organism. It is also interesting to observe that by taking nitrogen up as a mixture of nitrate and ammonium, plants apparently avoid the 'worst-case' nutrient efficiency modes, rather than selecting for the best. This bears a striking resemblance to some

EFM based genetic engineering strategies, as well as more diverse biological processes, such as natural search algorithms [85].

The potential existence of modes which are not biologically utilised means that although certain narratives, such as these, present themselves, care must be taken not to over-interpret the EFM distributions, and we do not claim that these hypotheses are the case. Obviously ideally some of these ideas would be further tested experimentally, fulfilling the mantra of iterative cycles of experiment, and modelling. Given the limited quality of the results seen under FBA analysis, in particular with regards to the comparison between sulfur starvation predictions, and transcriptomic data, it is not necessarily clear how accurate any of these ideas might prove to be. However, the modelling phase of 'idea generation' is complete.

## 5.4 Mobile mRNA

In spite of the above described broadening of scope, from sulfur assimilation, to nutrient use, we remained interested in sulfur metabolism, and spent some time investigating the bundle sheath expression of many genes associated with sulfur metabolism, and glucosinolate production. Thieme et al. [219] demonstrated that sulfur related genes were over represented amongst the mobile mRNA population, suggesting a potential reason for the observed bundle sheath gene expression pattern. In studying this paper, primarily due to this interest, we developed the results presented in chapter 4. Unexpectedly, this examination of the potential role of mRNA as long distance signalling molecules has proven to be probably the most successful area of study. Although it is clear that a few developmentally important examples exist of mRNA acting as a signalling molecule , we have shown that the potential of the large majority of previously reported mobile transcripts to be functionally significant remains unclear.

Given that similar results have since been reported with regards to a large scale study of protein mobility [159], it is an interesting, and I think, open question as to what extent phenomena which are observed, particularly in regards to 'omics studies are 1) functionally relevant, and 2) worthy of report. This raises questions as to the goals of fundamental biological science; is it to describe the biological process, or to understand why it is like that? In this example; is it interesting that transcripts are mobile if it makes no difference to the functioning of the plant?

It is easy to assume that everything observed which is not due to technical error is 'biologically important'; in particular, it seems common to assume that robustness to biological replication indicates that a result is relevant. However, as we have seen, given potential structure within 'error' terms, (in this case apparently caused by unrelated mRNA stability, and abundance), it is not clear that biological replicates provide a sufficient strategy for dealing with this

problem.

## 5.5 Conclusion

Analysis, particularly of EFMs, but also through FBA has demonstrated that relatively simple, constraint-based approaches can be usefully applied to hugely complex metabolic systems. They can be used to generate interesting hypotheses, in particular concerning the emergent properties of the reaction network as a system, I am here thinking particularly about flexibility, nutrient tradeoffs, and 'optimal' metabolic strategies. This is important because, as we have already touched upon in the discussion of FBA and transcriptomic response to sulfur starvation, it seems that people are predominantly interested in a phenomena if the cause, or motivation, for it can be explained. Consequently, many 'omics studies use stories to explain the 'purpose' behind their observations.

However, as we have seen in studying mobile mRNA, the presence of structured 'error' terms via functionally unconnected attributes can lead to the description of potentially non-functional phenomena, even though these phenomena can be narratively ascribed a potential purpose. It is perhaps ironic that parallels have emerged between the strengths of constraint-based approaches, in providing explanatory stories for the biological 'motivation' behind experimental observations of metabolism, and our work on mRNA; in which we have shown that the perhaps over enthusiastic assignation of narrative purpose has led beyond what the data justifies.

We should welcome the ability to generate more sophisticated, quantifiable stories about metabolic function, but we should also remember that although a particular story, or purpose presents itself, additional evidence, (beyond data driven observation), is generally required to demonstrate that this is in fact the case. A common position is to advocate a return to hypothesis driven experimentation, however it is not clear how this could itself resolve the problem. In the example considered, the hypothesis that 'mRNA is a common signalling molecule' would have been supported by the experimental data. This is clearly a hypothesis, but does not serve to overcome the difficulty. Consequently, it seems more relevant to emphasise the requirement of a higher standard of evidence, rather than a particular investigative paradigm. We therefore find that data-, rather than hypothesis-driven science is valuable, but it should be remembered that it is useful primarily for hypothesis generation, it is not itself convincing evidence for that hypothesis. Generally, further evidence is required. Hence we emphasise that the ideas of metabolism presented, particularly in chapter 3 require further experimental investigation.

In comparing predictions derived from this model, and experimental data, we have generally seen reasonable, although not strong agreement in terms of the responses of individual reactions to environmental perturbation, or the system to

perturbation of individual reactions. We conclude that, although genome scale models of plant metabolism continue to develop, and can already be usefully applied to specific areas of metabolism, they are not yet sufficiently sophisticated to be generally applicable.

Particular weaknesses are likely to be in the handling of subcellular compartmentalisation, but we have also seen that predictions are weaker in comparison to whole plants, or to organs than to cell suspension experiments. Consequently, it seems that the greatest stumbling block for these approaches currently is in the construction of decent plant models, accounting for tissue type, and intracellular transport. Although these models are extremely time consuming to develop, they have continuously advanced in quality. If, in particular, a unified, consensus model is adopted by the field, improvements will only increase in rapidity, and will likely advance to include high quality tissue specific models of whole plants. We have here demonstrated that EFM based methods are beginning to become computationally available for the analysis of genome scale models of plants, and hence methodological difficulties historically associated with FBA, and the idea of 'optimality' can potentially be sidestepped by sampling from the true metabolic space.

Although a somewhat winding path has been taken, this project has to date generated extensive advances in my personal understanding of, as well as some general insight into, Arabidopsis metabolism. More generally, it has provided the opportunity to develop from an out-and-out wet-lab scientist into something approaching a computational biologist, with experience in a number of programming languages, and a (statistically) unlikely interest in modelling.

# Bibliography


[1] **Allen DK** (2016) Quantifying plant phenotypes with isotopic labeling & metabolic flux analysis. Current Opinion in Biotechnology **37**:45–52.

[2] **Alves R, Vilaprinyo E, Hernandez-Bermejo B, Sorribas A** (2008) Mathematical formalisms based on approximated kinetic representations for modeling genetic and metabolic pathways. Biotechnology and Genetic Engineering Reviews 25.

[3] **Antoniewicz MR, Kelleher JK, Stephanopoulos G** (2007) Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. Metabolic Engineering **9**:68–86.

[4] **Apel K, Hirt H** (2004) Reactive oxygen species: Metabolism, oxidative stress, and signal transduction. Annual Review of Plant Biology **55**:373–399.

[5] **Arnold A, Nikoloski Z** (2014) Bottom-up metabolic reconstruction of arabidopsis and its application to determining the metabolic costs of enzyme production. Plant Physiology **165**(3):1380–1391.

[6] **Arnold A, Sajitz-Hermstein M, Nikoloski Z** (2015) Effects of varying nitrogen sources on amino acid synthesis costs in *Arabidopsis thaliana* under different light and carbon-source conditions. PLoS One **10**(2):e0116536.

[7] **Aubry S, Smith-Unna RD, Boursnell CM, Kopriva S, Hibberd JM** (2014) Transcript residency on ribosomes reveals a key role for the *Arabidopsis thaliana* bundle sheath in sulfur and glucosinolate metabolism. Plant Journal **78**(4):659–673.

[8] **Bailey TL** (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics **27**(12):1653–1659.

[9] **Banerjee AK, Chatterjee M, Yu Y, Suh SG, Miller WA, Hannapel DJ** (2006) Dynamics of a mobile RNA of potato involved in a long-distance signaling pathway. Plant Cell **18**(12):3443–3457.

[10] **Banerjee AK, Lin T, Hannapel DJ** (2009) Untranslated regions of a mobile transcript mediate RNA metabolism. Plant Physiology **151**(4):1831–1843.

[11] **Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D** (2007) How to infer gene networks from expression profiles. Molecular Systems Biology **3**:78.

[12] **Barrett CL, Herrgard MJ, Palsson B** (2009) Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation. BMC Systems Biology **3**:30.

[13] **Barve A, Rodrigues JFM, Wagner A** (2012) Superessential reactions in metabolic networks. Proceedings of the National Academy of Sciences of the United States of America **109**(18):E1121–E1130.

[14] **Becker SA, Palsson BO** (2008) Context-specific metabolic networks are consistent with experiments. PLoS Computational Biology **4**(5):e1000082.

[15] **Beer C, Reichstein M, Tomelleri E, Ciais P, Jung M, Carvalhais N, Rodenbeck C, Arain MA, Baldocchi D, Bonan GB, Bondeau A, Cescatti A, Lasslop G, Lindroth A, Lomas M, Luyssaert S, Margolis H, Oleson KW, Roupsard O, Veenendaal E, Viovy N, Williams C, Woodward FI, Papale D** (2010) Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. Science **329**(5993):834–838.

[16] **Behre J, Wilhelm T, von Kamp A, Ruppin E, Schuster S** (2008) Structural robustness of metabolic networks with respect to multiple knockouts. Journal of Theoretical Biology **252**(3):433–441.

[17] **Bekaert M, Edger PP, Hudson CM, Pires JC, Conant GC** (2012) Metabolic and evolutionary costs of herbivory defense: systems biology of glucosinolate synthesis. New Phytologist **196**(2):596–605.

[18] **Beurton-Aimar M, Beauvoit B, Monier A, Vallee F, Dieuaide-Noubhani M, Colombie S** (2011) Comparison between elementary flux modes analysis and 13C-metabolic fluxes measured in bacterial and plant cells. BMC Systems Biology **5**:95.

[19] **Birke H, Mueller SJ, Rother M, Zimmer AD, Hoernstein SNW, Wesenberg D, Wirtz M, Krauss GJ, Reski R, Hell R** (2012) The relevance of compartmentation for cysteine synthesis in phototrophic organisms. Protoplasma **249**:147–155.

[20] **Bloom AJ, Meyerhoff PA, Taylor AR, Rost TL** (2002) Root development and absorption of ammonium and nitrate from the rhizosphere. Journal of Plant Growth Regulation **21**(4):416–431.

[21] **Bogart E, Myers CR** (2016) Multiscale metabolic modeling of C4 plants: Connecting nonlinear genome-scale models to leaf-scale metabolism in developing maize leaves. PLoS One **11**(3):e0151722.

[22] **Bordbar A, Monk JM, King ZA, Palsson BO** (2014) Constraint-based models predict metabolic and associated cellular functions. Nature Reviews Genetics **15**(2):107–120.

[23] **Bortesi L, Fischer R** (2015) The CRISPR/Cas9 system for plant genome editing and beyond. Biotechnology Advances **33**(1):41–52.

[24] **Britto DT, Siddiqi MY, Glass ADM, Kronzucker HJ** (2001) Futile transmembrane $NH_4^+$ cycling: A cellular hypothesis to explain ammonium toxicity in plants. Proceedings of the National Academy of Sciences of the United States of America **98**(7):4255–4258.

[25] **Cao MJ, Wang Z, Wirtz M, Hell R, Oliver DJ, Xiang CB** (2013) SULTR3;1 is a chloroplast-localized sulfate transporter in *Arabidopsis thaliana*. Plant Journal **73**(4):607–616.

[26] **Carrari F, Urbanczyk-Wochniak E, Willmitzer L, Fernie AR** (2003) Engineering central metabolism in crop species: learning the system. Metabolic Engineering **5**:191–200.

[27] **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**(9):1043–1057.

[28] **Chan EKF, Rowe HC, Kliebenstein DJ** (2010) Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. Genetics **185**(3):991–1007.

[29] **Chen X, Alonso AP, Shachar-Hill Y** (2013) Dynamic metabolic flux analysis of plant cell wall synthesis. Metabolic Engineering **18**:78–85.

[30] **Chen X, Shachar-Hill Y** (2012) Insights into metabolic efficiency from flux analysis. Journal of Experimental Botany **63**(6):2343–2351.

[31] **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 C-3 and crassulacean acid metabolism leaves. Plant Physiology **165**(2):917–929.

[32] **Cheung CYM, Ratcliffe RG, Sweetlove LJ** (2015) A method of accounting for enzyme costs in flux balance analysis reveals alternative pathways and metabolite stores in an illuminated Arabidopsis leaf. Plant Physiology **169**(3):1671–1682.

[33] **Cheung CYM, Williams TCR, Poolman MG, Fell DA, Ratcliffe RG, Sweetlove LJ** (2013) A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell

metabolic phenotypes under stress conditions. Plant Journal **75**(6):1050–1061.

[34] **Collins SB, Reznik E, Segre D** (2012) Temporal expression-based analysis of metabolism. PLoS Computational Biology **8**(11):e1002781.

[35] **Colombie S, Nazaret C, Benard C, Biais B, Mengin V, Sole M, Fouillen L, Dieuaide-Noubhani M, Mazat JP, Beauvoit B, Gibon Y** (2015) Modelling central metabolic fluxes by constraint-based optimization reveals metabolic reprogramming of developing *Solanum lycopersicum* (tomato) fruit. Plant Journal **81**(1):24–39.

[36] **Copeland WB, Bartley BA, Chandran D, Galdzicki M, Kim KH, Sleight SC, Maranas CD, Sauro HM** (2012) Computational tools for metabolic engineering. Metabolic Engineering **14**:270–280.

[37] **Covert MW, Schilling CH, Palsson B** (2001) Regulation of gene expression in flux balance models of metabolism. Journal of Theoretical Biology **213**(1):73–88.

[38] **Covert MW, Xiao N, Chen TJ, Karr JR** (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. Bioinformatics **24**(18):2044–2050.

[39] **Curien G, Cardenas ML, Cornish-Bowden A** (2014) Analytical kinetic modeling: a practical procedure. Methods in Molecular Biology (Clifton, NJ) **1090**:261–280.

[40] **Curien G, Ravanel S, Dumas R** (2003) A kinetic model of the branch-point between the methionine and threonine biosynthesis pathways in *Arabidopsis thaliana*. European Journal of Biochemistry **270**(23):4615–4627.

[41] **Dal'Molin CGdO, Nielsen LK** (2013) Plant genome-scale metabolic reconstruction and modelling. Current Opinion in Biotechnology **24**(2):271–277.

[42] **Dal'Molin CGdO, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK** (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. Plant Physiology **152**(2):579–589.

[43] **Dal'Molin CGdO, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK** (2010) C4GEM, a genome-scale metabolic model to study C-4 plant metabolism. Plant Physiology **154**(4):1871–1885.

[44] **Dal'Molin CGdO, Quek LE, Saa PA, Nielsen LK** (2015) A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. Frontiers in Plant Science **6**:4.

[45] **Dantzig GB, Orden A, Wolfe P, et al.** (1955) The generalized simplex method for minimizing a linear form under linear inequality restraints. Pacific Journal of Mathematics **5**(2):183–195.

[46] **David L, Bockmayr A** (2014) Computing elementary flux modes involving a set of target reactions. IEEE/ACM Transactions on Computational Biology and Bioinformatics **11**(6):1099–1107.

[47] **de Figueiredo LF, Podhorski A, Rubio A, Kaleta C, Beasley JE, Schuster S, Planes FJ** (2009) Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics **25**(23):3158–3165.

[48] **D'Hooghe P, Escamez S, Trouverie J, Avice JC** (2013) Sulphur limitation provokes physiological and leaf proteome changes in oilseed rape that lead to perturbation of sulphur, carbon and oxidative metabolisms. BMC Plant Biology **13**:23.

[49] **Dias O, Rocha M, Ferreira EC, Rocha I** (2015) Reconstructing genome-scale metabolic models with MERLIN. Nucleic Acids Research **43**(8):3899–3910.

[50] **Du JL, Yuan ZF, Ma ZW, Song JZ, Xie XL, Chen YL** (2014) KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Molecular Biosystems **10**(9):2441–2447.

[51] **Durot M, Bourguignon PY, Schachter V** (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiology Reviews **33**(1):164–190.

[52] **Estevez SR, Nikoloski Z** (2015) Context-specific metabolic model extraction based on regularized least squares optimization. PLoS One **10**(7):e0131875.

[53] **Famili I, Mahadevan R, Palsson BO** (2005) K-cone analysis: Determining all candidate values for kinetic parameters on a network scale. Biophysical Journal **88**(3):1616–1625.

[54] **Fang K, Zhao H, Sun C, Lam CMC, Chang S, Zhang K, Panda G, Godinho M, dos Santos VAPM, Wang J** (2011) Exploring the metabolic network of the epidemic pathogen *Burkholderia cenocepacia* J2315 via genome-scale reconstruction. BMC Systems Biology **5**:83.

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

[56] **Feist AM, Palsson BO** (2010) The biomass objective function. Current Opinion in Microbiology **13**(3):344–349.

[57] **Fischer E, Sauer U** (2005) Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. Nature Genetics **37**(6):636–640.

[58] **Fisher AK, Freedman BG, Bevan DR, Senger RS** (2014) A review of metabolic and enzymatic engineering strategies for designing and optimizing performance of microbial cell factories. Computational and Structural Biotechnology Journal **11**(18):91–99.

[59] **Fisher CP, Plant NJ, Moore JB, Kierzek AM** (2013) QSSPN: dynamic simulation of molecular interaction networks describing gene regulation, signalling and whole-cell metabolism in human cells. Bioinformatics **29**(24):3181–3190.

[60] **Fusco D, Accornero N, Lavoie B, Shenoy SM, Blanchard JM, Singer RH, Bertrand E** (2003) Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. Current Biology **13**(2):161–167.

[61] **Gagneur J, Klamt S** (2004) Computation of elementary modes: a unifying framework and the new binary approach. BMC Bioinformatics **5**:1.

[62] **Garcia-Contreras R, Vos P, Westerhoff HV, Boogerd FC** (2012) Why in vivo may not equal in vitro - new effectors revealed by measurement of enzymatic activities under the same in vivo-like assay conditions. FEBS Journal **279**(22):4145–4159.

[63] **Gauthier PPG, Bligny R, Gout E, Mahe A, Nogues S, Hodges M, Tcherkez GGB** (2010) In folio isotopic tracing demonstrates that nitrogen assimilation into glutamate is mostly independent from current $CO_2$ assimilation in illuminated leaves of *Brassica napus*. New Phytologist **185**(4):988–999.

[64] **Gerstl MP, Jungreuthmayer C, Muller S, Zanghellini J** (2016) Which sets of elementary flux modes form thermodynamically feasible flux distributions? FEBS Journal **283**(9):1782–1794.

[65] **Gerstl MP, Klamt S, Jungreuthmayer C, Zanghellini J** (2016) Exact quantification of cellular robustness in genome-scale metabolic networks. Bioinformatics **32**(5):730–737.

[66] **Gerstl MP, Ruckerbauer DE, Mattanovich D, Jungreuthmayer C, Zanghellini J** (2015) Metabolomics integrated elementary flux mode analysis in large metabolic networks. Scientific Reports **5**:8930.

[67] **Gilroy S, Suzuki N, Miller G, Choi WG, Toyota M, Devireddy AR, Mittler R** (2014) A tidal wave of signals: calcium and ROS at the forefront of rapid systemic signaling. Trends in Plant Science **19**(10):623–630.

[68] **Glaeser K, Kanawati B, Kubo T, Schmitt-Kopplin P, Grill E** (2014) Exploring the Arabidopsis sulfur metabolome. Plant Journal **77**(1):31–45.

[69] **Goel A, Santos F, Vos WM, Teusink B, Molenaar D** (2012) Standardized assay medium to measure *Lactococcus lactis* enzyme activities while mimicking intracellular conditions. Applied and Environmental Microbiology **78**(1):134–143.

[70] **Graf A, Schlereth A, Stitt M, Smith AM** (2010) Circadian control of carbohydrate availability for growth in Arabidopsis plants at night. Proceedings of the National Academy of Sciences of the United States of America **107**(20):9458–9463.

[71] **Grafahrend-Belau E, Junker A, Eschenroeder A, Mueller J, Schreiber F, Junker BH** (2013) Multiscale metabolic modeling: Dynamic flux balance analysis on a whole-plant scale. Plant Physiology **163**(2):637–647.

[72] **Grafahrend-Belau E, Schreiber F, Heiner M, Sackmann A, Junker BH, Grunwald S, Speer A, Winder K, Koch I** (2008) Modularization of biochemical networks based on classification of Petri net t-invariants. BMC Bioinformatics **9**:90.

[73] **Grafahrend-Belau E, Schreiber F, Koschuetzki D, Junker BH** (2009) Flux balance analysis of barley seeds: A computational approach to study systemic properties of central metabolism. Plant Physiology **149**(1):585–598.

[74] **Gu BJ, Chang J, Min Y, Ge Y, Zhu QA, Galloway JN, Peng CH** (2013) The role of industrial nitrogen in the global nitrogen biogeochemical cycle. Scientific Reports **3**:2579.

[75] **Hachiya T, Watanabe CK, Fujimoto M, Ishikawa T, Takahara K, Kawai-Yamada M, Uchimiya H, Uesono Y, Terashima I, Noguchi K** (2012) Nitrate addition alleviates ammonium toxicity without lessening ammonium accumulation, organic acid depletion and inorganic cation depletion in *Arabidopsis thaliana* shoots. Plant and Cell Physiology **53**(3):577–591.

[76] **Hadicke O, Klamt S** (2011) Computing complex metabolic intervention strategies using constrained minimal cut sets. Metabolic Engineering **13**:204–213.

[77] **Haedicke O, Klamt S** (2010) CASOP: A computational approach for strain optimization aiming at high productivity. Journal of Biotechnology **147**(2):88–101.

[78] **Harcombe WR, Delaney NF, Leiby N, Klitgord N, Marx CJ** (2013) The ability of flux balance analysis to predict evolution of cen-

tral metabolism scales with the initial distance to the optimum. PLoS Computational Biology **9**(6):e1003091.

[79] **Hastie T, Tibshirani R, Friedman J** (2001) The Elements of Statistical Learning. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[80] **Hawkesford M** (2014) Nutrient use efficiency in plants : concepts and approaches. Springer, Cham.

[81] **Hawkins DM** (2004) The problem of overfitting. Journal of Chemical Information and Computer Sciences **44**:1.

[82] **Hay J, Schwender J** (2011) Computational analysis of storage synthesis in developing *brassica napus* l. (oilseed rape) embryos: flux variability analysis in relation to C-13 metabolic flux analysis. Plant Journal **67**(3):513–525.

[83] **Hay JO, Shi H, Heinzel N, Hebbelmann I, Rolletschek H, Schwender J** (2014) Integration of a constraint-based metabolic model of *Brassica napus* developing seeds with C-13 metabolic flux analysis. Frontiers in Plant Science **5**:724.

[84] **Haywood V, Yu TS, Huang NC, Lucas WJ** (2005) Phloem long-distance trafficking of gibberellic acid-insensitive RNA regulates leaf development. Plant Journal **42**(1):49–68.

[85] **Hein AM, Carrara F, Brumley DR, Stocker R, Levin SA** (2016) Natural search algorithms as a bridge between organisms, evolution, and ecology. Proceedings of the National Academy of Sciences of the United States of America **113**(34):9413–9420.

[86] **Hernandez I, Munne-Bosch S** (2015) Linking phosphorus availability with photo-oxidative stress in plants. Journal of Experimental Botany **66**(10):2889–2900.

[87] **Hirai MY, Fujiwara T, Awazuhara M, Kimura T, Noji M, Saito K** (2003) Global expression profiling of sulfur-starved Arabidopsis by DNA macroarray reveals the role of O-acetyl-L-serine as a general regulator of gene expression in response to sulfur nutrition. Plant Journal **33**(4):651–663.

[88] **Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K, Goda H, Nishizawa OI, Shibata D, Saito K** (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proceedings of the National Academy of Sciences of the United States of America **104**(15):6478–6483.

[89] **Horvat P, Koller M, Braunegg G** (2015) Recent advances in elementary flux modes and yield space analysis as useful tools in metabolic net-

work studies. World Journal of Microbiology & Biotechnology **31**(9):1315–1328.

[90] **Hosseini Z, Marashi SA** (2015) Hierarchical organization of fluxes in *Escherichia coli* metabolic network: Using flux coupling analysis for understanding the physiological properties of metabolic genes. Gene **561**(2):199–208.

[91] **Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang MZ, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY** (2001) The Arabidopsis information resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Research **29**(1):102–105.

[92] **Huang DW, Sherman BT, Lempicki RA** (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols **4**(1):44–57.

[93] **Hunt KA, Folsom JP, Taffs RL, Carlson RP** (2014) Complete enumeration of elementary flux modes through scalable demand-based subnetwork definition. Bioinformatics **30**(11):1569–1578.

[94] **Ibarra RU, Edwards JS, Palsson BO** (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420**(6912):186–189.

[95] **Ip K, Colijn C, Lun DS** (2011) Analysis of complex metabolic behavior through pathway decomposition. BMC Systems Biology **5**:91.

[96] **Janga SC** (2012) From specific to global analysis of posttranscriptional regulation in eukaryotes: posttranscriptional regulatory networks. Briefings in Functional Genomics **11**(6):505–521.

[97] **Jazmin LJ, O'Grady JP, Ma F, Allen DK, Morgan JA, Young JD** (2014) Isotopically nonstationary MFA (INST-MFA) of autotrophic metabolism. Methods in Molecular Biology **1090**:181–210.

[98] **Jones AM, Danielson JAH, ManojKumar SN, Lanquar V, Grossmann G, Frommer WB** (2014) Abscisic acid dynamics in roots detected with genetically encoded FRET sensors. eLife **3**:e01741.

[99] **Jungreuthmayer C, Ruckerbauer DE, Zanghellini J** (2013) regEfmtool: speeding up elementary flux mode calculation using transcriptional regulatory rules in the form of three-state logic. Biosystems **113**(1):37–39.

[100] **Kaleta C, de Figueiredo LF, Behre J, Schuster S** (2009) EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. Lecture Notes in Informatics-Proceedings **157**:179–189.

[101] **Kaltenbach HM, Stelling J** (2012) Modular Analysis of Biological Networks, volume 736 of *Advances in Experimental Medicine and Biology*, 3–17.

[102] **Karp PD, Paley S, Romero P** (2002) The pathway tools software. Bioinformatics **18**(Suppl 1):S225–S232.

[103] **Kehr J, Buhtz A** (2008) Long distance transport and movement of RNA through the phloem. Journal of Experimental Botany **59**(1):85–92.

[104] **Kelk SM, Olivier BG, Stougie L, Bruggeman FJ** (2012) Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. Scientific Reports **2**:580.

[105] **Kesler SE** (2007) Mineral supply and demand into the 21st century. In Proceedings for a Workshop on Deposit Modeling, Mineral Resource Assessment, and Their Role in Sustainable Development., volume 1294, 55–62. U.S. Geological Survey.

[106] **Khodayari A, Zomorrodi AR, Liao JC, Maranas CD** (2014) A kinetic model of *Escherichia coli* core metabolism satisfying multiple sets of mutant flux data. Metabolic Engineering **25**:50–62.

[107] **Kim G, LeBlanc ML, Wafula EK, dePamphilis CW, Westwood JH** (2014) Genomic-scale exchange of mRNA between a parasitic plant and its hosts. Science **345**(6198):808–811.

[108] **Kim M, Canio W, Kessler S, Sinha N** (2001) Developmental changes due to long-distance movement of a homeobox fusion transcript in tomato. Science **293**(5528):287–289.

[109] **Kitano H** (2010) Violations of robustness trade-offs. Molecular Systems Biology **6**:384.

[110] **Klamt S, Stelling J** (2002) Combinatorial complexity of pathway analysis in metabolic networks. Molecular Biology Reports **29**(1-2):233–236.

[111] **Koohkan H, Maftoun M** (2016) Effect of nitrogen-boron interaction on plant growth and tissue nutrient concentration of canola (*Brassica napus* L.). Journal of Plant Nutrition **39**(7):922–931.

[112] **Kopriva S, Rennenberg H** (2004) Control of sulphate assimilation and glutathione synthesis: interaction with N and C metabolism. Journal of Experimental Botany **55**(404):1831–1842.

[113] **Koprivova A, Suter M, Op den Camp R, Brunold C, Kopriva S** (2000) Regulation of sulfate assimilation by nitrogen in Arabidopsis. Plant Physiology **122**(3):737–746.

[114] **Krauss M, Schaller S, Borchers S, Findeisen R, Lippert J, Kuepfer L** (2012) Integrating cellular metabolism into a multiscale whole-body model. PLoS Computational Biology **8**(10):e1002750.

[115] **Krueger S, Giavalisco P, Krall L, Steinhauser MC, Buessis D, Usadel B, Fluegge UI, Fernie AR, Willmitzer L, Steinhauser D** (2011) A topological map of the compartmentalized *Arabidopsis thaliana* leaf metabolome. PLoS One **6**(3):e17806.

[116] **Kumar A, Harrelson T, Lewis NE, Gallagher EJ, LeRoith D, Shiloach J, Betenbaugh MJ** (2014) Multi-tissue computational modeling analyzes pathophysiology of type 2 diabetes in MKR mice. PLoS One **9**(7):e102319.

[117] **Kutz A, Muller A, Hennig P, Kaiser WM, Piotrowski M, Weiler EW** (2002) A role for Nitrilase 3 in the regulation of root morphology in sulphur-starving *Arabidopsis thaliana*. Plant Journal **30**(1):95–106.

[118] **Labhsetwar P, Cole JA, Roberts E, Price ND, Luthey-Schulten ZA** (2013) Heterogeneity in protein expression induces metabolic variability in a modeled *Escherichia coli* population. Proceedings of the National Academy of Sciences of the United States of America **110**(34):14006–14011.

[119] **Lam HM, Coschigano K, Schultz C, Melooliveira R, Tjaden G, Oliveira I, Ngai N, Hsieh MH, Coruzzi G** (1995) Use of Arabidopsis mutants and genes to study amide amino-acid biosynthesis. Plant Cell **7**(7):887–898.

[120] **Laporte DC, Walsh K, Koshland DE** (1984) The branch point effect - ultrasensitivity and subsensitivity to metabolic control. Journal of Biological Chemistry **259**(22):4068–4075.

[121] **Larhlimi A, Blachon S, Selbig J, Nikoloski Z** (2011) Robustness of metabolic networks: A review of existing definitions. Biosystems **106**:1.

[122] **Leighty RW, Antoniewicz MR** (2011) Dynamic metabolic flux analysis (DMFA): a framework for determining fluxes at metabolic non-steady state. Metabolic Engineering **13**:745–755.

[123] **Leighty RW, Antoniewicz MR** (2013) COMPLETE-MFA: complementary parallel labeling experiments technique for metabolic flux analysis. Metabolic Engineering **20**:49–55.

[124] **Leroux AE, Haanstra JR, Bakker BM, Krauth-Siegel RL** (2013) Dissecting the catalytic mechanism of *Trypanosoma brucei* trypanothione synthetase by kinetic analysis and computational modeling. Journal of Biological Chemistry **288**(33):23751–23764.

[125] **Lewis NE, Nagarajan H, Palsson BO** (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nature Reviews Microbiology **10**(4):291–305.

[126] **Liang D, White RG, Waterhouse PM** (2012) Gene silencing in Arabidopsis spreads from the root to the shoot, through a gating barrier, by

template-dependent, nonvascular, cell-to-cell movement. Plant Physiology **159**(3):984–1000.

[127] **Link H, Christodoulou D, Sauer U** (2014) Advancing metabolic models with kinetic information. Current Opinion in Biotechnology **29C**:8–14.

[128] **Lloyd J, Meinke D** (2012) A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. Plant Physiology **158**(3):1115–1129.

[129] **Lough TJ, Lucas WJ** (2006) Integrative plant biology: Role of phloem long-distance macromolecular trafficking. Annual Review of Plant Biology **57**:203–232.

[130] **Machado D, Herrgard M** (2014) Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Computational Biology **10**(4):e1003580.

[131] **Machado D, Soons Z, Patil KR, Ferreira EC, Rocha I** (2012) Random sampling of elementary flux modes in large-scale metabolic networks. Bioinformatics **28**(18):i515–i521.

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

[133] **Mahajan A, Bhogale S, Kang IH, Hannapel DJ, Banerjee AK** (2012) The mRNA of a knotted1-like transcription factor of potato is phloem mobile. Plant Molecular Biology **79**(6):595–608.

[134] **Martins Conde PdR, Sauter T, Pfau T** (2016) Constraint based modelling going multicellular. Frontiers in Molecular Biosciences **3**:3.

[135] **Maruyama-Nakashita A, Inoue E, Watanabe-Takahashi A, Yarnaya T, Takahashi H** (2003) Transcriptome profiling of sulfur-responsive genes in Arabidopsis reveals global effects of sulfur nutrition on multiple metabolic pathways. Plant Physiology **132**(2):597–605.

[136] **Maruyama-Nakashita A, Nakamura Y, Tohge T, Saito K, Takahashi H** (2006) Arabidopsis SLIM1 is a central transcriptional regulator of plant sulfur response and metabolism. Plant Cell **18**(11):3235–3251.

[137] **Masakapalli SK, Bryant FM, Kruger NJ, Ratcliffe RG** (2014) The metabolic flux phenotype of heterotrophic Arabidopsis cells reveals a flexible balance between the cytosolic and plastidic contributions to carbohydrate oxidation in response to phosphate limitation. Plant Journal **78**(6):964–977.

[138] **Masakapalli SK, Kruger NJ, Ratcliffe RG** (2013) The metabolic flux phenotype of heterotrophic Arabidopsis cells reveals a complex response to changes in nitrogen supply. Plant Journal **74**(4):569–582.

[139] **Masakapalli SK, Le Lay P, Huddleston JE, Pollock NL, Kruger NJ, Ratcliffe RG** (2010) Subcellular flux analysis of central metabolism in a heterotrophic arabidopsis cell suspension using steady-state stable isotope labeling. Plant Physiology **152**(2):602–619.

[140] **Melzer G, Esfandabadi ME, Franco-Lara E, Wittmann C** (2009) Flux design: In silico design of cell factories based on correlation of pathway fluxes to desired properties. BMC Systems Biology **3**:120.

[141] **Mendoza-Cozatl DG, Moreno-Sanchez R** (2006) Control of glutathione and phytochelatin synthesis under cadmium stress. pathway modeling for plants. Journal of Theoretical Biology **238**(4):919–936.

[142] **Michael TP, McClung CR** (2003) Enhancer trapping reveals widespread circadian clock transcriptional control in Arabidopsis. Plant Physiology **132**(2):629–639.

[143] **Mignone F, Gissi C, Liuni S, Pesole G** (2002) Untranslated regions of mRNAs. Genome Biology **3**:3.

[144] **Min Y, Jin XG, Chen M, Pan ZZ, Ge Y, Chang J** (2011) Pathway knockout and redundancy in metabolic networks. Journal of Theoretical Biology **270**(1):63–69.

[145] **Mintz-Oron S, Meir S, Malitsky S, Ruppin E, Aharoni A, Shlomi T** (2012) Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. Proceedings of the National Academy of Sciences of the United States of America **109**(1):339–344.

[146] **Miskovic L, Hatzimanikatis V** (2011) Modeling of uncertainties in biochemical reactions. Biotechnology and Bioengineering **108**(2):413–423.

[147] **Mustroph A, Zanetti ME, Jang CJH, Holtan HE, Repetti PP, Galbraith DW, Girke T, Bailey-Serres J** (2009) Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **106**(44):18843–18848.

[148] **Narsai R, Howell KA, Millar AH, O'Toole N, Small I, Whelan J** (2007) Genome-wide analysis of mRNA decay rates and their determinants in *Arabidopsis thaliana*. Plant Cell **19**(11):3418–3436.

[149] **Nero D, Krouk G, Tranchina D, Coruzzi GM** (2009) A system biology approach highlights a hormonal enhancer effect on regulation of genes in a nitrate responsive "biomodule". BMC Systems Biology **3**:59.

[150] **Newman DJ, Cragg GM** (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. Journal of Natural Products **75**(3):311–335.

[151] **Nikiforova V, Freitag J, Kempa S, Adamik M, Hesse H, Hoefgen R** (2003) Transcriptome analysis of sulfur depletion in *Arabidopsis thaliana*: interlacing of biosynthetic pathways provides response specificity. Plant Journal **33**(4):633–650.

[152] **Nikiforova VJ, Kopka J, Tolstikov V, Fiehn O, Hopkins L, Hawkesford MJ, Hesse H, Hoefgen R** (2005) Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of Arabidopsis plants. Plant Physiology **138**(1):304–318.

[153] **Nookaew I, Meechai A, Thammarongtham C, Laoteng K, Ruanglek V, Cheevadhanarak S, Nielsen J, Bhumiratana S** (2007) Identification of flux regulation coefficients from elementary flux modes: A systems biology tool for analysis of metabolic networks. Biotechnology and Bioengineering **97**(6):1535–1549.

[154] **Notaguchi M, Higashiyama T, Suzuki T** (2015) Identification of mRNAs that move over long distances using an RNA-Seq analysis of Arabidopsis/*Nicotiana benthamiana* heterografts. Plant and Cell Physiology **56**(2):311–321.

[155] **Notaguchi M, Wolf S, Lucas WJ** (2012) Phloem-mobile Aux/IAA transcripts target to the root tip and modify root architecture. Journal of Integrative Plant Biology **54**(10):760–772.

[156] **Oberhardt MA, Palsson BO, Papin JA** (2009) Applications of genome-scale metabolic reconstructions. Molecular Systems Biology **5**:320.

[157] **Papp B, Pal C, Hurst LD** (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature **429**(6992):661–664.

[158] **Patterson K, Cakmak T, Cooper A, Lager I, Rasmusson AG, Escobar MA** (2010) Distinct signalling pathways and transcriptome response signatures differentiate ammonium- and nitrate-supplied plants. Plant Cell and Environment **33**(9):1486–1501.

[159] **Paultre DS, Gustin MP, Molnar A, Oparka KJ** (2016) Lost in transit: long-distance trafficking and phloem unloading of protein signals in Arabidopsis homografts. Plant Cell **28**:2016–2025.

[160] **Peres S, Vallee F, Beurton-Aimar M, Mazat JP** (2011) ACoM: A classification method for elementary flux modes based on motif finding. Biosystems **103**(3):410–419.

[161] **Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, McMurry J, Marioni JC, Malone J, Megy K, Rustici G, Tang AY, Taubert J, Williams E, Mannion O, Parkinson HE, Brazma A**

(2014) Expression Atlas update-a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Research **42**(D1):D926–D932.

[162] **Pey J, Planes FJ** (2014) Direct calculation of elementary flux modes satisfying several biological constraints in genome-scale metabolic networks. Bioinformatics **30**(15):2197–2203.

[163] **Pey J, Villar JA, Tobalina L, Rezola A, Manuel Garcia J, Beasley JE, Planes FJ** (2015) TreeEFM: calculating elementary flux modes using linear optimization in a tree-based algorithm. Bioinformatics **31**(6):897–904.

[164] **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**(7):1673–1682.

[165] **Pilkington SM, Encke B, Krohn N, Hoehne M, Stitt M, Pyl ET** (2015) Relationship between starch degradation and carbon demand for maintenance and growth in *Arabidopsis thaliana* in different irradiance and temperature regimes. Plant Cell and Environment **38**(1):157–171.

[166] **Ponzoni I, Nueda MJ, Tarazona S, Gotz S, Montaner D, Dussaut JS, Dopazo J, Conesa A** (2014) Pathway network inference from gene expression data. BMC Systems Biology **8**(Suppl 2):S7.

[167] **Poolman MG, Fell DA, Raines CA** (2003) Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. European Journal of Biochemistry **270**(3):430–439.

[168] **Poolman MG, Kundu S, Shaw R, Fell DA** (2014) Metabolic tradeoffs between biomass synthesis and photosynthate export at different light intensities in a genome-scale metabolic model of rice. Frontiers in Plant Science **5**:656.

[169] **Poolman MG, Miguet L, Sweetlove LJ, Fell DA** (2009) A genome-scale metabolic model of Arabidopsis and some of its properties. Plant Physiology **151**(3):1570–1581.

[170] **Poolman MG, Sebu C, Pldcock MK, Fell DA** (2007) Modular decomposition of metabolic systems via null-space analysis. Journal of Theoretical Biology **249**(4):691–705.

[171] **Poolman MG, Venkatesh KV, Pidcock MK, Fell DA** (2004) A method for the determination of flux in elementary modes, and its application to *Lactobacillus rhamnosus*. Biotechnology and Bioengineering **88**(5):601–612.

[172] **Pozo C, Miro A, Guillen-Gosalbez G, Sorribas A, Alves R, Jimenez L** (2015) Gobal optimization of hybrid kinetic/FBA models via outer-approximation. Computers & Chemical Engineering **72**:325–333.

[173] **Price ND, Papin JA, Palsson BO** (2002) Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. Genome Research **12**(5):760–769.

[174] **Price ND, Reed JL, Palsson BO** (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nature Reviews Microbiology **2**(11):886–897.

[175] **Reimers AC** (2015) Hierarchical decomposition of metabolic networks using k-modules. Biochemical Society Transactions **43**:1146–1150.

[176] **Rezola A, Pey J, de Figueiredo LF, Podhorski A, Schuster S, Rubio A, Planes FJ** (2013) Selection of human tissue-specific elementary flux modes using gene expression data. Bioinformatics **29**(16):2009–2016.

[177] **Rohwer JM** (2014) Applications of kinetic modeling to plant metabolism. Methods in Molecular Biology (Clifton, NJ) **1083**:275–86.

[178] **Rohwer JM, Botha FC** (2001) Analysis of sucrose accumulation in the sugar cane culm on the basis of in vitro kinetic data. Biochemical Journal **358**:437–445.

[179] **Roman MS, Cancela H, Acerenza L** (2014) Source and regulation of flux variability in *Escherichia coli*. BMC Systems Biology **8**:67.

[180] **Roosta HR, Schjoerring JK** (2008) Effects of nitrate and potassium on ammonium toxicity in cucumber plants. Journal of Plant Nutrition **31**(7):1270–1283.

[181] **Sackmann A, Heiner M, Koch I** (2006) Application of Petri net based analysis techniques to signal transduction pathways. BMC Bioinformatics **7**:482.

[182] **Saha R, Chowdhury A, Maranas CD** (2014) Recent advances in the reconstruction of metabolic models and integration of omics data. Current Opinion in Biotechnology **29**(0):39–45.

[183] **Sajitz-Hermstein M, Nikoloski Z** (2010) A novel approach for determining environment-specific protein costs: the case of *Arabidopsis thaliana*. Bioinformatics **26**(18):i582–i588.

[184] **Scheerer U, Haensch R, Mendel RR, Kopriva S, Rennenberg H, Herschbach C** (2010) Sulphur flux through the sulphate assimilation pathway is differently controlled by adenosine 5'-phosphosulphate reductase under stress and in transgenic poplar plants overexpressing gamma-ECS, SO, or APR. Journal of Experimental Botany **61**(2):609–622.

[185] **Schilling CH, Letscher D, Palsson BO** (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. Journal of Theoretical Biology **203**(3):229–248.

[186] **Schmidt BJ, Ebrahim A, Metz TO, Adkins JN, Palsson BO, Hyduke DR** (2013) GIM(3)E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics **29**(22):2900–2908.

[187] **Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U** (2012) Multidimensional optimality of microbial metabolism. Science **336**(6081):601–604.

[188] **Schultz A, Qutub AA** (2016) Reconstruction of tissue-specific metabolic networks using CORDA. PLoS Computational Biology **12**(3):e1004808.

[189] **Schuster S, Fell DA, Dandekar T** (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature Biotechnology **18**(3):326–332.

[190] **Schuster S, Hilgetag C** (1994) On elementary flux modes in biochemical reaction systems at steady state. Journal of Biological Systems **2**(2):165–182.

[191] **Schuster S, Klamt S, Weckwerth W, Moldenhauer F, Pfeiffer T** (2002) Use of network analysis of metabolic systems in bioengineering. Bioprocesses and Biosystems Engineering **24**:363–372.

[192] **Schuster S, Pfeiffer T, Fell DA** (2008) Is maximization of molar yield in metabolic networks favoured by evolution? Journal of Theoretical Biology **252**(3):497–504.

[193] **Schwender J, Goffman F, Ohlrogge JB, Shachar-Hill Y** (2004) Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature **432**(7018):779–782.

[194] **Schwender J, Koenig C, Klapperstueck M, Heinzel N, Munz E, Hebbelmann I, Hay JO, Denolf P, De Bodt S, Redestig H, Caestecker E, Jakob PM, Borisjuk L, Rolletschek H** (2014) Transcript abundance on its own cannot be used to infer fluxes in central metabolism. Frontiers in Plant Science **5**:668.

[195] **Searles PS, Bloom AJ** (2003) Nitrate photo-assimilation in tomato leaves under short-term exposure to elevated carbon dioxide and low oxygen. Plant Cell and Environment **26**(8):1247–1255.

[196] **Seaver SMD, Henry CS, Hanson AD** (2012) Frontiers in metabolic reconstruction and modeling of plant genomes. Journal of Experimental Botany **63**(6):2247–2258.

[197] **Segre D, Vitkup D, Church GM** (2002) Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences of the United States of America **99**(23):15112–15117.

[198] **Shah J, Zeier J** (2013) Long-distance communication and signal amplification in systemic acquired resistance. Frontiers in Plant Science **4**:30.

[199] **Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T** (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research **13**(11):2498–2504.

[200] **Shlomi T, Berkman O, Ruppin E** (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proceedings of the National Academy of Sciences of the United States of America **102**(21):7695–7700.

[201] **Simons M, Misra A, Sriram G** (2014) Genome-scale models of plant metabolism. Methods in Molecular Biology (Clifton, NJ) **1083**:213–230.

[202] **Simons M, Saha R, Amiour N, Kumar A, Guillard L, Clement G, Miquel M, Li Z, Mouille G, Lea PJ, Hirel B, Maranas CD** (2014) Assessing the metabolic impact of nitrogen availability using a compartmentalized maize leaf genome-scale model. Plant Physiology **166**(3):1659–1674.

[203] **Singer SD, Zou J, Weselake RJ** (2016) Abiotic factors influence plant storage lipid accumulation and composition. Plant Science **243**:1–9.

[204] **Smallbone K, Mendes P** (2013) Large-scale metabolic models: from reconstruction to differential equations. Industrial Biotechnology **9**(4):179–184.

[205] **Smallbone K, Simeonidis E** (2009) Flux balance analysis: a geometric perspective. Journal of Theoretical Biology **258**(2):311–315.

[206] **Sonderby IE, Geu-Flores F, Halkier BA** (2010) Biosynthesis of glucosinolates - gene discovery and beyond. Trends in Plant Science **15**(5):283–290.

[207] **Song HS, Ramkrishna D** (2009) Reduction of a set of elementary modes using yield analysis. Biotechnology and Bioengineering **102**(2):554–568.

[208] **Sparks E, Wachsman G, Benfey PN** (2013) Spatiotemporal signalling in plant development. Nature Reviews Genetics **14**(9):631–644.

[209] **Spiegelman Z, Golan G, Wolf S** (2013) Don't kill the messenger: long-distance trafficking of mRNA molecules. Plant Science **213**:1–8.

[210] **Srienc F, Unrean P** (2010) A statistical thermodynamical interpretation of metabolism. Entropy **12**(8):1921–1935.

[211] **Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED** (2002) Metabolic network structure determines key aspects of functionality and regulation. Nature **420**(6912):190–193.

[212] **Steuer R, Gross T, Selbig J, Blasius B** (2006) Structural kinetic modeling of metabolic networks. Proceedings of the National Academy of Sciences of the United States of America **103**(32):11868–11873.

[213] **Stitt M, Gibon Y** (2014) Why measure enzyme activities in the era of systems biology? Trends in Plant Science **19**(4):256–265.

[214] **Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, Zielinski DC, Ang KS, Gardiner NJ, Gutierrez JM, Kyriakopoulos S, Lakshmanan M, Li S, Liu JK, Martinez VS, Orellana CA, Quek LE, Thomas A, Zanghellini J, Borth N, Lee DY, Nielsen LK, Kell DB, Lewis NE, Mendes P** (2016) Recon 2.2: from reconstruction to model of human metabolism. Metabolomics **12**:109.

[215] **Szecowka M, Heise R, Tohge T, Nunes-Nesi A, Vosloh D, Huege J, Feil R, Lunn J, Nikoloski Z, Stitt M, Fernie AR, Arrivault S** (2013) Metabolic fluxes in an illuminated Arabidopsis rosette. Plant Cell **25**(2):694–714.

[216] **Takahashi H, Kopriva S, Giordano M, Saito K, Hell R** (2011) Sulfur assimilation in photosynthetic organisms: molecular functions and regulation of transporters and assimilatory enzymes, volume 62 of *Annual Review of Plant Biology*, 157–184.

[217] **Terzer M, Stelling J** (2008) Large-scale computation of elementary flux modes with bit pattern trees. Bioinformatics **24**(19):2229–2235.

[218] **Thiele I, Palsson BO** (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols **5**(1):93–121.

[219] **Thieme CJ, Rojas-Triana M, Stecyk E, Schudoma C, Zhang W, Yang L, Miambres M, Walther D, Schulze WX, Paz-Ares J, Scheible WR, Kragler F** (2015) Endogenous Arabidopsis messenger RNAs transported to distant tissues. Nature Plants **1**:15025.

[220] **Toepfer N, Caldana C, Grimbs S, Willmitzer L, Fernie AR, Nikoloski Z** (2013) Integration of genome-scale modeling and transcript profiling reveals metabolic pathways underlying light and temperature acclimation in Arabidopsis. Plant Cell **25**(4):1197–1211.

[221] **Traka M, Mithen R** (2009) Glucosinolates, isothiocyanates and human health. Phytochemistry Reviews **8**:269 – 282.

[222] **Tran LM, Rizk ML, Liao JC** (2008) Ensemble modeling of metabolic networks. Biophysical Journal **95**(12):5606–5617.

[223] **Trinh CT, Carlson R, Wlaschin A, Srienc F** (2006) Design, construction and performance of the most efficient biomass producing *E. coli* bacterium. Metabolic Engineering **8**:628–638.

[224] **Trinh CT, Wlaschin A, Srienc F** (2009) Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Applied Microbiology and Biotechnology **81**(5):813–826.

[225] **Tschoep H, Gibon Y, Carillo P, Armengaud P, Szecowka M, Nunes-Nesi A, Fernie AR, Koehl K, Stitt M** (2009) Adjustment of growth and central metabolism to a mild but sustained nitrogen-limitation in Arabidopsis. Plant Cell and Environment **32**(3):300–318.

[226] **Tummler K, Lubitz T, Schelker M, Klipp E** (2014) New types of experimental data shape the use of enzyme kinetics for dynamic network modeling. FEBS Journal **281**(2):549–571.

[227] **Urbanczik R, Wagner C** (2005) Functional stoichiometric analysis of metabolic networks. Bioinformatics **21**(22):4176–4180.

[228] **van Eunen K, Bakker BM** (2014) The importance and challenges of in vivo-like enzyme kinetics. Perspectives in Science **1**(1):126–130.

[229] **van Klinken JB, Willems van Dijk K** (2016) FluxModeCalculator: an efficient tool for large-scale flux mode computation. Bioinformatics **32**(8):1265–1266.

[230] **Varma A, Palsson BO** (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. Applied and Environmental Microbiology **60**(10):3724–3731.

[231] **Vauclare P, Kopriva S, Fell D, Suter M, Sticher L, von Ballmoos P, Krahenbuhl U, den Camp RO, Brunold C** (2002) Flux control of sulphate assimilation in *Arabidopsis thaliana*: adenosine 5 '-phosphosulphate reductase is more susceptible than ATP sulphurylase to negative control by thiols. Plant Journal **31**(6):729–740.

[232] **Vig AP, Rampal G, Thind TS, Arora S** (2009) Bio-protective effects of glucosinolates - a review. LWT-Food Science and Technology **42**(10):1561–1572.

[233] **Vivek-Ananth RP, Samal A** (2016) Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. Biosystems 147.

[234] **von Kamp A, Klamt S** (2014) Enumeration of smallest intervention strategies in genome-scale metabolic networks. PLoS Computational Biology **10**(1):e1003378.

[235] **von Kamp A, Schuster S** (2006) Metatool 5.0: fast and flexible elementary modes analysis. Bioinformatics **22**(15):1930–1931.

[236] **von Stosch M, de Azevedo CR, Luis M, de Azevedo SF, Oliveira R** (2016) A principal components method constrained by elementary flux modes: analysis of flux data sets. BMC Bioinformatics **17**:200.

[237] **Wagner C** (2004) Nullspace approach to determine the elementary modes of chemical reaction systems. Journal of Physical Chemistry B **108**(7):2425–2431.

[238] **Wang C, Guo L, Li Y, Wang Z** (2012) Systematic comparison of C3 and C4 plants based on metabolic network analysis. BMC Systems Biology **6**(Suppl 2):S9.

[239] **Wang Q, Yang Y, Ma H, Zhao X** (2007) Metabolic network properties help assign weights to elementary modes to understand physiological flux distributions. Bioinformatics **23**(9):1049–1052.

[240] **Watanabe M, Hubberten HM, Saito K, Hoefgen R** (2010) General regulatory patterns of plant mineral nutrient depletion as revealed by serat quadruple mutants disturbed in cysteine synthesis. Molecular Plant **3**(2):438–466.

[241] **Westwood JH** (2015) RNA transport: Delivering the message. Nature Plants **1**:15038.

[242] **Wilhelm T, Behre J, Schuster S** (2004) Analysis of structural robustness of metabolic networks. Systems Biology (Stevenage) **1**(1):114–120.

[243] **Williams TCR, Miguet L, Masakapalli SK, Kruger NJ, Sweetlove LJ, Ratcliffe RG** (2008) Metabolic network fluxes in heterotrophic Arabidopsis cells: Stability of the flux distribution under different oxygenation conditions. Plant Physiology **148**(2):704–718.

[244] **Williams TCR, Poolman MG, Howden AJM, Schwarzlander M, Fell DA, Ratcliffe RG, Sweetlove LJ** (2010) A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. Plant Physiology **154**(1):311–323.

[245] **Wintermute EH, Lieberman TD, Silver PA** (2013) An objective function exploiting suboptimal solutions in metabolic networks. BMC Systems Biology **7**:98.

[246] **Witte JS** (2010) Genome-wide association studies and beyond. Annual Review of Public Health **31**:9–20.

[247] **Yoon J, Si Y, Nolan R, Lee K** (2007) Modular decomposition of metabolic reaction networks based on flux analysis and pathway projection. Bioinformatics **23**(18):2433–2440.

[248] **Yoshimoto N, Inoue E, Watanabe-Takahashi A, Saito K, Takahashi H** (2007) Posttranscriptional regulation of high-affinity sulfate transporters in Arabidopsis by sulfur nutrition. Plant Physiology **145**(2):378–388.

[249] **Yuan H, Cheung CYM, Hilbers PAJ, van Riel NAW** (2016) Flux balance analysis of plant metabolism: The effect of biomass composition and model structure on model predictions. Frontiers in Plant Science **7**:537.

[250] **Yuan L, Grotewold E** (2015) Metabolic engineering to enhance the value of plants as green factories. Metabolic Engineering **27**:83–91.

[251] **Zamora-Sillero E, Hafner M, Ibig A, Stelling J, Wagner A** (2011) Efficient characterization of high-dimensional parameter spaces for systems biology. BMC Systems Biology **5**:142.

[252] **Zarecki R, Oberhardt MA, Yizhak K, Wagner A, Segal ES, Freilich S, Henry CS, Gophna U, Ruppin E** (2014) Maximal sum of metabolic exchange fluxes outperforms biomass yield as a predictor of growth rate of microorganisms. PLoS One **9**(5):e0098372.

[253] **Zhang PF, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY** (2005) MetaCyc and AraCyc. metabolic pathway databases for plant research. Plant Physiology **138**(1):27–37.

[254] **Zhang W, Thieme CJ, Kollwig G, Apelt F, Yang L, Winter N, Andresen N, Walther D, Kragler F** (2016) tRNA-related sequences trigger systemic mRNA transport in plants. Plant Cell **28**(6):1237–1249.

[255] **Zuker A, Tzfira T, Ben-Meir H, Ovadis M, Shklarman E, Itzhaki H, Forkmann G, Martens S, Neta-Sharir I, Weiss D, Vainstein A** (2002) Modification of flower color and fragrance by antisense suppression of the flavanone 3-hydroxylase gene. Molecular Breeding **9**(1):33–41.