Data driven approaches to infer the regulatory mechanism shaping and constraining levels of metabolites in metabolic networks

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium
in der Wissenschaftsdisziplin “Systembiologie”

eingereicht in kumulativer Form
an der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam von

Kevin Schwahn

Disputation am 20.12.2018

Betreuer:

Prof. Dr. Alisdair R. Fernie
Prof. Dr. Zoran Nikoloski

Gutachter:

Prof. Dr. Zoran Nikoloski
Dr. Joachim Kopka
Prof. Dr. Björn Usadel

Published online at the
Institutional Repository of the University of Potsdam:
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-423240
https://doi.org/10.25932/publishup-42324

Contents

1 Abstract

Systems biology aims at investigating biological systems in its entirety by gathering and analyzing large-scale data sets about the underlying components. Computational systems biology approaches use these large-scale data sets to create models at different scales and cellular levels. In addition, it is concerned with generating and testing hypotheses about biological processes. However, such approaches are inevitably leading to computational challenges due to the high dimensionality of the data and the differences in the dimension of data from different cellular layers.

This thesis focuses on the investigation and development of computational approaches to analyze metabolite profiles in the context of cellular networks. This leads to determining what aspects of the network functionality are reflected in the metabolite levels. With these methods at hand, this thesis aims to answer three questions: (1) how observability of biological systems is manifested in metabolite profiles and if it can be used for phenotypical comparisons; (2) how to identify couplings of reaction rates from metabolic profiles alone; and (3) which regulatory mechanism that affect metabolite levels can be distinguished by integrating transcriptomics and metabolomics read-outs.

I showed that sensor metabolites, identified by an approach from observability theory, are more correlated to each other than non-sensors. The greater correlations between sensor metabolites were detected both with publicly available metabolite profiles and synthetic data simulated from a medium-scale kinetic model. I demonstrated through robustness analysis that correlation was due to the position of the sensor metabolites in the network and persisted irrespectively of the experimental conditions. Sensor metabolites are therefore potential candidates for phenotypical comparisons between conditions through targeted metabolic analysis.

Furthermore, I demonstrated that the coupling of metabolic reaction rates can be investigated from a purely data-driven perspective, assuming that metabolic reactions can be described by mass action kinetics. Employing metabolite profiles from domesticated and wild wheat and tomato species, I showed that the process of domestication is associated with a loss of regulatory control on the level of reaction rate coupling. I also found that the same metabolic pathways in *Arabidopsis thaliana* and *Escherichia coli* exhibit differences in the number of reaction rate couplings.

I designed a novel method for the identification and categorization of transcriptional effects on metabolism by combining data on gene expression and metabolite levels. The approach determines the partial correlation of metabolites with control by the principal components of the transcript levels. The principle components contain the majority of the transcriptomic information allowing to partial out the effect of the transcriptional layer from the metabolite profiles. Depending whether the correlation between metabolites persists upon controlling for the effect of the transcriptional layer, the approach allows us to group metabolite pairs into being associated due to post-transcriptional or transcriptional regulation, respectively. I showed that the classification of metabolite pairs into those that are associated due to transcriptional or post-transcriptional regulation are in agreement with existing literature and findings from a Bayesian inference approach.

The approaches developed, implemented, and investigated in this thesis open novel ways to jointly study metabolomics and transcriptomics data as well as to place metabolic profiles in the network context. The results from these approaches have the potential to provide further insights into the regulatory machinery in a biological system.

2 Zusammenfassung

Die System Biologie ist auf die Auswertung biologischer Systeme in ihrer Gesamtheit gerichtet. Dies geschieht durch das Sammeln und analysieren von großen Datensätzen der zugrundeliegenden Komponenten der Systeme. Computergestützte systembiologische Ansätze verwenden diese großen Datensätze, um Modelle zu erstellen und Hypothesen über biologische Prozesse auf verschiedenen zellularen Ebenen zu testen. Diese Ansätze führen jedoch unweigerlich zu rechnerischen Herausforderungen, da die Daten über eine hohe Dimensionalität verfügen. Des Weiteren weisen Daten, die von verschiedenen zellulären Ebenen gewonnen werden, unterschiedliche Dimensionen auf.

Diese Doktorarbeit beschäftigt sich mit der Untersuchung und Entwicklung von rechnergestützten Ansätzen, um Metabolit-Profile im Zusammenhang von zellulären Netzwerken zu analysieren und um zu bestimmen, welche Aspekte der Netzwerkfunktionalität sich in den Metabolit-Messungen widerspiegeln. Die Zielsetzung dieser Arbeit ist es, die folgenden Fragen, unter Berücksichtigung der genannten Methoden, zu beantworten: (1) Wie ist die Beobachtbarkeit von biologischen Systemen in Metabolit-Profilen manifestiert und sind diese für phänotypische Vergleiche verwendbar? (2) Wie lässt sich die Kopplung von Reaktionsraten ausschließlich durch Metabolit-Profile identifizieren? (3) Welche regulatorischen Mechanismen, die Metabolit-Niveaus beeinflussen, sind unterscheidbar, wenn transkriptomische und metabolische Daten kombiniert werden?

Ich konnte darlegen, dass Sensormetabolite, die durch eine Methode der „observability theory“ identifiziert wurden, stärker korrelieren als Nicht-Sensoren. Die stärkere Korrelation zwischen Sensormetaboliten konnte mit öffentlich zugänglichen Daten, als auch mit synthetischen Daten aus einer Simulation mit einem mittelgroßen kinetischen Modell gezeigt werden. Durch eine Robustheitsanalyse war es mir möglich zu demonstrieren, dass die Korrelation auf die Position der Sensormetabolite im Netzwerk zurückzuführen und unabhängig von den experimentellen Bedingungen ist. Sensormetabolite sind daher geeignete Kandidaten für phänotypische Vergleiche zwischen verschiedenen Bedingungen durch gezielte metabolische Analysen.

Des Weiteren ergaben meine Untersuchungen, dass die Auswertung der Kopplung von Stoffwechselreaktionsraten von einer ausschließlich datengestützten Perspektive möglich ist. Dabei muss die Annahme getroffen werden, dass Stoffwechselreaktionen mit dem Massenwirkungsgesetz beschreibbar sind. Ich konnte zeigen, dass der Züchtungsprozess mit einem Verlust der regulatorischen Kontrolle auf der Ebene der gekoppelten Reaktionsraten einhergeht. Dazu verwendete ich Metabolit-Profile von gezüchteten, als auch wilden Weizen- und Tomatenspezies. Meine Ergebnisse belegen, dass die selben Stoffwechselwege in *Arabidopsis thaliana und Escherichia coli* eine unterschiedliche Anzahl an gekoppelten Reaktionsraten aufweisen.

Darüber hinaus habe ich eine neue Methode zur Identifizierung und Kategorisierung von transkriptionellen Effekten auf den Metabolismus entwickelt. Dies erfolgt durch die Kombination von Genexpressionsdaten und Messungen von Metaboliten. Die Methode ermittelt die partielle Korrelation zwischen Metaboliten, wobei die Hauptkomponenten der Transkriptdaten als Kontrollvariablen dienen. Dadurch kann der Einfluss der Transkription auf Metabolit-Profile herausgerechnet werden. Dieser Ansatz ermöglicht die Einteilung von Metabolitpaaren in assoziiert durch transkriptionelle oder assoziiert durch posttranskriptionelle Regulation. Die Einteilung ist abhängig davon, ob die Korrelation zwischen Metaboliten bestehen bleibt, wenn für den Einfluss der Transkription kontrolliert wird. Ich

konnte nachweisen, dass die zuvor genannten Klassifizierungen von Metabolitpaaren mit existierender Literatur und den Ergebnissen einer auf bayessche Statistik basierenden Studie übereinstimmen.

Die Methoden, die in dieser Doktorarbeit entwickelt, implementiert und untersucht wurden, öffnen neue Wege um metabolische und transkriptomische Daten gemeinsam auszuwerten. Sie erlauben Metabolit-Profile in den Kontext von metabolischen Netzwerken zu stellen. Die Ergebnisse haben das Potential uns weitere Einblicke in die regulatorische Maschinerie in biologischen Systemen zu gewähren.

3 Introduction

During the 20th century, biology focused on the investigation of specific cellular components, their function and localization. While these investigations have provided insights into the function of fundamental cellular components, they have largely neglected the connections among the components and the resulting mutual dependence of cellular processes. Systems biology emerged over the last two decades with a focus on investigating entire systems instead of single biological components (e.g. genes or proteins). The advent of systems biology is linked to the availability of data sets generated by high-throughput technologies giving rise to the *omics* fields including: genomics [Campos-de Quiroz, 2002], epigenomics [Köhler and Springer, 2017], transcriptomics [Usadel and Fernie, 2013], proteomics [Baginsky, 2009] and metabolomics [Fiehn, 2002]. The most impactful technological advances as a result were genome sequencing, RNA-microarrays, RNA-sequencing, mass spectrometry (MS) and Nuclear Magnetic Resonance (NMR)-technologies. The availability of these high-throughput data has propelled biologists to shift their focus from analyzing single components to investigating and understanding entire cells and organisms as complex systems.

Multiple large-scale approaches can either be used for top-down or bottom-up systems biology analysis. In top-down approaches one or more large-scale data sets are used to investigate a biological process in question and for testing posited hypotheses [Chuang et al., 2010]. In contrast, bottom-up systems biology focuses on the creation of models at different scales from those involving few interacting components to genome-scale cellular networks. These models are generated from annotations of genome sequences and manual curation. The bottom-up approaches can be employed to test for missing reactions in the constructed network, the effect of gene knockouts or the comparison of phenotypic observations with predictions from simulations [Heavner and Price, 2015; Benedict et al., 2012]. Therefore, holistic questions in the context of biological systems can be answered with the combination of multiple data sets originating from different *omics* technologies and models of different complexity. It is then apparent that advances in systems biology necessitate the design, implementation and testing of reliable methods to facilitate the top-down and bottom-up approaches.

In the context of top-down approaches, the investigation of large-scale data sets in general imposes some challenges. First, one needs to take into account the high dimensionality of the data itself. Transcriptomic technologies can be used to measure several thousand of transcripts yielding complete overview over the transcribed genes [Jain, 2012]. Multiple gene expression data sets can be combined to generate co-expression networks by calculating similarity scores between the genes over multiple conditions and data sets [Serin et al., 2016]. In contrast, around thousand metabolites can be monitored per experiment which is only a portion of the metabolome [Vinaixa et al., 2012; Giavalisco et al., 2008]. This already indicates that the difference in dimensions has to be taken into account as well when comparing data gathered at different cellular layers. Additionally, in most biological studies the number of observations (time points and conditions) are lower than the number of measured genes or metabolites [Caldana et al., 2011; Jozefczuk et al., 2010]. This high number of variables (p) in comparison to the number of observations (n) leads to a "large p, small n" problem for both metabolomic and transcriptomic data sets. That leads to numerical and computational issues in many classical approaches, such as regression methods, as parameters in the regression model can not be reliably estimated [Johnstone and Titterington, 2009]. The problem of high dimensionality can be addressed by performing dimension reduction or employing regularization methods [Adragni and Cook, 2009;

Johnstone and Titterington, 2009]. The investigation of metabolite levels alone or combined with transcriptomic measurements requires statistical approaches that take these requirements into account.

This thesis is comprised of three published studies that fall under the umbrella of systems biology. These studies focus on design, implementation, and application of novel methods to investigate levels of metabolites and their regulation, primarily in the context of plant science. In the first two I only investigated metabolite levels relating them on the structure of metabolic networks; while in the third I investigated novel approaches for the combined investigation of metabolite and transcript levels. The introduced approaches are purely data-driven and are based on correlation. These approaches will increase the amount of information that can be retrieved from already performed experiments and will allow a deeper understanding of regulatory processes in biological systems.

3.1 Metabolism and the means to measure its constituting components

The metabolome is the entirety of all metabolites within an organism [Fiehn, 2002]. Metabolites fulfill a wide range of functions and have acquired a broad range of chemical properties [Hartmann, 2007]. They can be classified into the two categories of primary and secondary metabolites: Primary metabolites are essential for maintenance and growth of the organism. In contrast, secondary metabolites cover a broad range of secondary functions, such as: defense and stress tolerance [Hartmann, 2007]. In every cell metabolites are chemically transformed by special proteins called enzymes. Multiple consecutive enzyme-catalyzed reactions can be combined into metabolic pathways and further represented in metabolic networks. A genome-scale metabolic network contains a complete overview of the known metabolic reactions of an organism [de Oliveira Dal'Molin et al., 2010a]. Figure 3.1 visualizes the differentiation of a single metabolic pathway in comparison to a metabolic network. In sub-figure 3.1A, the glycolysis is presented: a metabolic pathway converting glucose to pyruvate while forming the high-energy metabolites ATP and NADH [Voet and Voet, 2011]. In contrast sub-figure 3.1B, shows the central metabolism of *Escherichia coli*, a sub-part of the genome-scale metabolic network [Orth et al., 2011]. The network structure highly affects the functions of metabolic pathways regulated a multitude of mechanisms on the level of transcription, translation and post-transcriptional modifications of enzymes. Measurements of metabolites therefore capture the combined outcome of these regulatory effects. This makes metabolite studies ideal for the identification of environmental induced changes and add another level for phenotypic comparison of biological systems.

3.1.1 Network representation of metabolites and metabolism

Metabolic networks represented in figure 3.1A mainly serves the purpose of visualization. While this representation is understandable to a human, it does not allow for the modeling of cellular dynamics. This can be achieved by representing the system of interest with ordinary differential equations (ODEs). The ODEs are derived from laws of mass balance and describe the change of a metabolite (x_j) over time, taking into account its production and utilization (see Figure 3.2A/B) [Hageman Blair et al., 2012; Schwender and Junker, 2009].

Equation 3.1 describes the change of metabolite x_j as the sum product of the fluxesv_i and the stoichiometric coefficient α_{ij} with which the metabolite enters the i^{th} reaction.

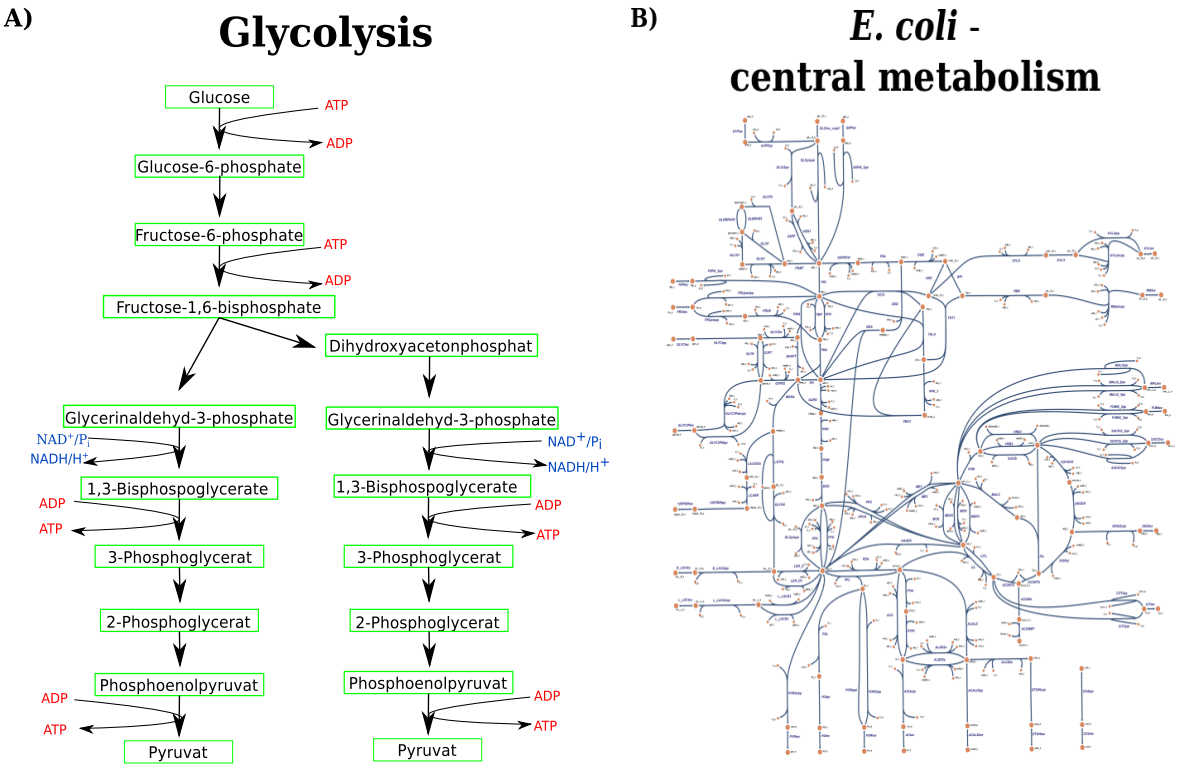

Figure 3.1: **Example pathway and metabolic network.**
A) Detailed representation of the glycolysis pathway and its metabolites. Shown in red are ATP/ADP and in blue the involved reducing-equivalents $NAD^+/NADH/H^+$. The glycolysis is a central pathway and converts one molecule of glucose into two molecules of pyruvate, while generating two molecules of $NADH/H^+$. **B)** Illustration of a the central metabolism from the *Escherichia coli* genome-scale metabolic model iJO1366 [Orth et al., 2011], drawn with Escher [King et al., 2015]. Nodes represent metabolites in the network, whereas the edges represent the reaction connecting the metabolites.

$$\frac{dx_j}{dt} = \sum_{i=1}^{n} \alpha_{ij} v_i \tag{3.1}$$

The systems of equations of a system can be represented in a matrix-vector form:

$$\frac{dx}{dt} = S * v \tag{3.2}$$

where S is the stoichiometric matrix of size $m \times n$, with m being the number of metabolites and n the number of reactions and v the vector of all fluxes $v_i, 1 \leq i \leq n$. Positive entries in the stoichiometric matrix corresponds to the production and each negative entry corresponds to the consumption of a metabolite. The representation of metabolic networks based on equation 3.2 allows to either simulate metabolite concentrations with kinetic modeling or metabolite reaction fluxes with stoichiometric modeling [Schwender and Junker, 2009].

Kinetic modeling can be performed with mass action kinetics to model enzymatic reactions. In mass action kinetics, the reaction velocity is proportional to the concentration of the substrates of the reaction and is formulated as:

$$v_i = k_i \prod_j x_j^{\alpha_{ij}}, \tag{3.3}$$

where v_i is the reaction rate, k_i is the rate constant, x_j the concentration of the substrate and α_{ij} is the stoichiometry with which metabolite x_j enters reaction i [Voit et al., 2015]. This estimation allows to describe the dynamics of metabolomic networks with few parameters, as only the metabolite concentration and the rate constant are needed, but all reactions should be elementary. The mass action kinetic can be extended towards the Michaelis-Menten kinetic requiring additional enzymatic parameters. The Michaelis-Menten kinetic is often associated with the quasi-steady-state assumptions assuming that during the conversion of a substrate S into the product P, the intermediate enzyme-substrate (ES) complex concentration does not change over time. In general, a metabolic reaction is assumed to follow this reaction scheme:

$$S + E \underset{k_{-1}}{\overset{k_1}{\longleftrightarrow}} ES \xrightarrow{k_2} E + P$$

.

Based on the above reaction formulation, a set of ODEs can be formulated and analytically solved, which results in a formulation of the reaction rate:

$$v = k_2 E_{total} \frac{S}{K_M + S} = V_{max} \frac{S}{K_m + S}, \tag{3.4}$$

where $k_2 E_{total} = V_{max}$ is the maximum reaction velocity with E_{total} stands for the total enzyme concentration, S the substrate concentration and $\frac{k_{-1}+k_2}{k_1} = K_M$ the Michaelis-Menten constant [Michaelis and Menten, 1913; Schallau and Junker, 2010]. A small K_M indicates high affinity of the enzyme for the substrate, implying a rate closer to $V_{\max}$. The value of K_M is dependent on both the enzyme and the substrate as well as experimental conditions. Therefore, the construction of a kinetic model based

on Michaelis-Menten kinetics requires detailed knowledge about the enzyme specific parameters K_M and V_{max} as well as metabolite and enzyme concentrations (Figure 3.2C). However, these parameters are often not available as experimental evaluations and conditions are missing or only available from *in vitro* measurements. This might not be truly indicative of the *in vivo* conditions [Schwender and Junker, 2009].

In contrast to kinetic modeling, stoichiometric modeling can be performed without the knowledge of kinetic parameters and relies only on the stoichiometry of the metabolites within the reaction network. The approach assumes that the system is in a steady-state, such that the concentration of intracellular metabolites does not change over time:

$$\frac{dx}{dt} = S * v = 0 \tag{3.5}$$

Therefore, a system of linear algebraic equations can be used to solve for the flux vector v, which is easier than using a system of ordinary differential equations [Schwender and Junker, 2009]. An approach that purely relies on the stoichiometry of the network is flux balance analysis (FBA) [Orth et al., 2010]. FBA seeks to maximize or minimize an objective function (Z) which in most cases represent the growth of the organism and is formulated as a linear program:

$$\begin{aligned} max(min)\colon Z &= c^t v \\ Sv &= 0 \\ v_L \leq 0 &\leq v_U \end{aligned} \tag{3.6}$$

Here, c is a vector of weights, indicating how much each reaction contributes to the objective function. Further, fluxes through the system can be constrained by v_U and v_L the upper and lower bounds (Figure 3.2D). The linear formulation of the problem allows to simulate metabolite fluxes for large networks.

3.1.2 Metabolomics technologies

Metabolites are believed to be closest to a system's phenotype thus capturing directly the response of the system to internal and external perturbations. Mass spectrometry (MS) and Nuclear Magnetic Resonance (NMR)-technologies are the most frequently used technologies to investigate the level of metabolites under different conditions, with MS-based approaches being predominantly used in plant studies [Fernie et al., 2004; Jorge et al., 2016b]. Metabolomics approaches can be divided into target and untargeted approaches [Johnson et al., 2016]. Targeted approaches are used to measure the levels for a set of given (known) metabolites. This requires external standards and calibration curves for each metabolite [Johnson et al., 2016; Lei et al., 2011]. In contrast, untargeted metabolomics allow to measure a wide range of metabolites present in the sample. In this approach relative quantification is performed which normalizes each metabolite signal to an internal standard. The internal standard is a metabolite that otherwise is not found in the sample [Jorge et al., 2016a; Lei et al., 2011], such as cholesterol [Jozefczuk et al., 2010] or ribitol [Lisec et al., 2006]. Further, it is possible to retain fold-changes of the metabolite content if a treatment and a control measurements were performed. The calculation of fold-changes is possible with absolute concentrations and relative metabolite levels [Vinaixa

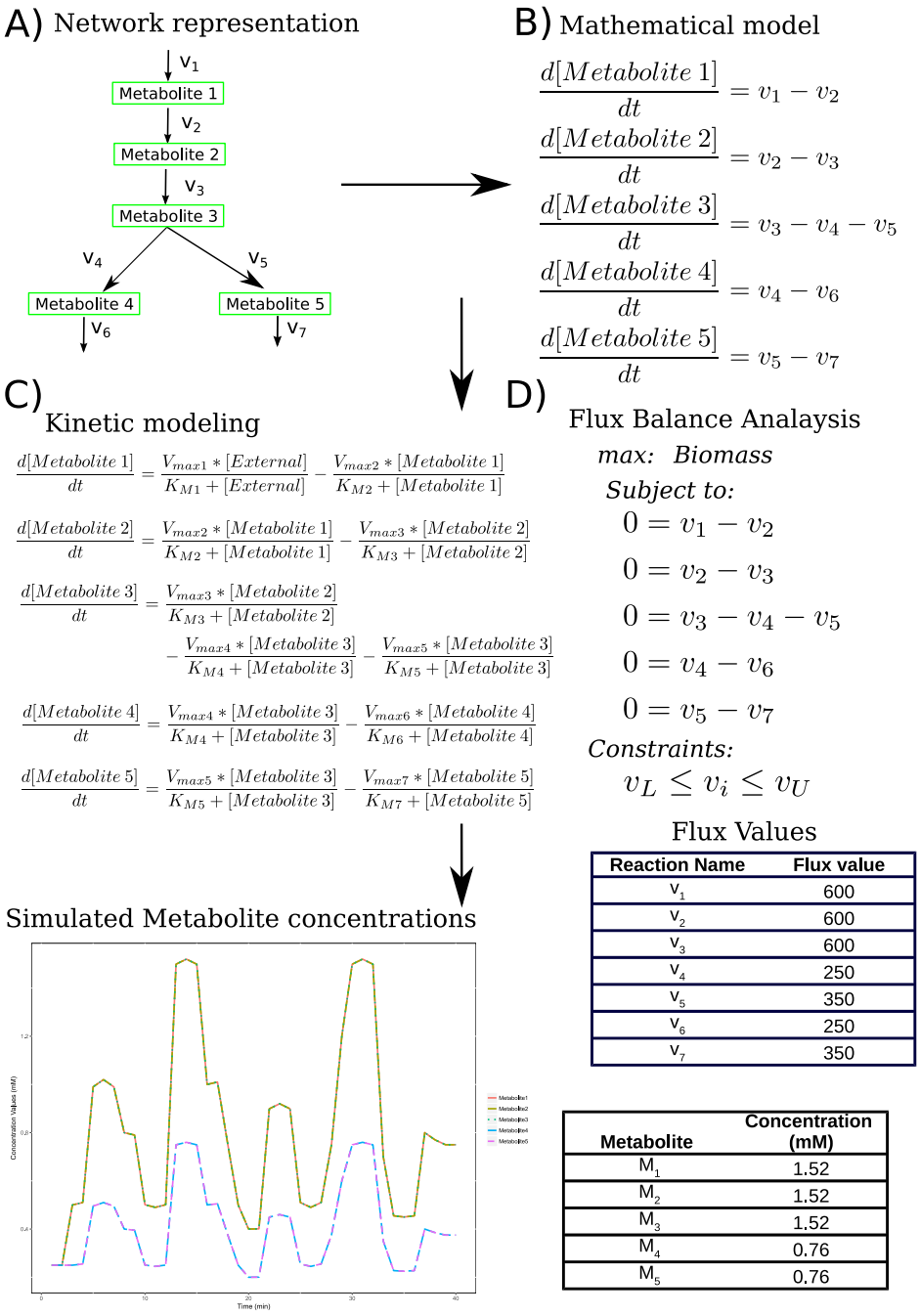

Figure 3.2: **Schematic overview of kinetic and stoichiometric modeling**
A) Example Metabolic network composed of five metabolites and seven reactions. **B)** The network can be described by a system of ordinary differential equations (ODEs), based on mass balance. **C)** The dynamics within the network can be analyzed with kinetic modeling. In the example, the reactions are modeled with Michaelis-Menten kinetics, which requires detailed knowledge about enzymatic parameters. Solving the system of differential equations results in a trajectory for each metabolite concentration. Intermediate concentrations are shown in a separate table. **D)** Alternatively, flux balance analysis (FBA) can be used, which assumes the system to be at a steady state. FBA solves a system of linear algebraic equations, which requires an objective function and additional constraints, resulting in a solution of fluxes corresponding to the steady state.

et al., 2012; Fernie et al., 2004]. Independent of this classification, measurements can either be done from experiments which capture or neglect the dynamics of the changes. Static measurements result in a snapshot of the state of the system whereas dynamic measurements allow us to obtain time-series data capturing the changes of the system [Antoniewicz, 2015; Peters et al., 2010].

Different separation and ionization techniques can be combined and are used for the investigation of the plant metabolome resulting in multiple combinations with MS methods. The general concept of most of the possible approaches is to separate the metabolites in the sample, to break them into smaller fragments and to introduce a charge through the ionization [Alonso et al., 2015; Jorge et al., 2016b]. The introduced charge is important for the acceleration of the fragments within the magnetic field and for the detection itself, as different fragments will have different mass-to-charge ratios (m/z). Subsequently, each metabolite will result in several peaks in the spectra corresponding to the different fragments [El-Aneed et al., 2009; Theodoridis et al., 2011].

From a technical point of view MS methods can be distinguished based on the separation technique. The most frequently used separation techniques are liquid chromatography (LC) and gas chromatography (GC). The use of a separation method increases the detection precision because in addition to the m/z value, the retention time adds a second value to distinguish between different fragments of the same m/z value. In principle, this technique relies on the enormous diversity of chemical properties of the metabolites which will influence their interaction with the adsorbent material and subsequently affect their elution time [Alonso et al., 2015]. Whether LC-MS or GC-MS needs to be applied depends largely on the type of metabolites to be investigated.

GC-MS has been primarily used to reliably investigate the primary metabolism, such as: sugars, organic acids and amino acids [Lisec et al., 2006; Caldana et al., 2011; Mönchgesang et al., 2016]. This approach is constrained by the volatility of the metabolites and is limited to those that are volatile (e.g. short chain alcohols, acid, esters, and hydrocarbons), or can be rendered volatile by derivatization [Lei et al., 2011]. During the derivatization processes the polar groups, such as N-H or O-H, of metabolites are converted into nonpolar groups rendering the metabolite more volatile. GC-MS technologies are frequently combined with electron ionization (EI) or chemical ionization (CI) techniques [Schwender and Junker, 2009]. The ionization of the metabolites with EI is performed through an electron beam that is produced by accelerated electrons moving toward a positive charged trap. The electrons will energize the metabolites. The energized metabolites seek a lower energy state by emitting an electron resulting in positive-charged molecules [Sparkman et al., 2011]. The CI technology is based on a similar principle whereby the charge to the metabolites is introduced through a reagent gas under high pressure by an electron beam. The reagent molecules are activated by the electrons from the electron beam and are ionized. Subsequently, they will then ionize the metabolites from the sample through collision [Sparkman et al., 2011].

In contrast to GC-MS, LC-MS based approaches allow to investigate metabolites in a wider range of chemical properties. Secondary metabolites, like flavonoids, hydroxycinnametes and steroidal glycoalkaloids, have been investigated in tomato fruits [Tohge et al., 2014; Moco et al., 2006], as well as glucosinolates, flavonoids and sterol lipids in *Arabidopsis thaliana* [Mönchgesang et al., 2016; Wewer et al., 2011]. LC-MS approaches are frequently combined with Electrospray Ionization (ESI), Atmospheric Pressure Chemical Ionization (APCI) or Matrix-Assisted Laser Desorption Ionization (MALDI) [Schwender and Junker, 2009]. The working principle of ESI is that the metabolite sample flows at a steady rate through a spray needle at a high voltage. The voltage, being either positive or negative, is applied to the sample and introduces the required charge on the metabolite fragments [Awad et al.,

2015]. ESI can be used with a wide range of metabolites and can generate multiple fragments for larger molecules [El-Aneed et al., 2009; Awad et al., 2015]. In APCI, the sample is sprayed under atmospheric pressure through a vaporization chamber in which the solvent is vaporized. Next, the molecules pass a discharge electrode where the metabolites are ionized. In comparison to ESI, APCI has an improved performance at higher flow rates [Awad et al., 2015]. Metabolite analysis by MALDI requires that the molecules and a solvent are spotted on a plate together with a matrix and dried. A laser beam then excites the matrix, transferring the energy on the molecules, which causes desorption and ionization of the molecules [Fernie et al., 2004; El-Aneed et al., 2009; Awad et al., 2015].

After the metabolites have been separated based on their chemical properties and their resulting mass fragments have been ionized, further separation is performed based on the m/z value in the mass analyzer. GC-MS, as well as LC-MS, can be combined with various mass analyzers of which the quadrupole mass analyzer is a frequently used example [Jorge et al., 2016b]. A quadrupole mass analyzer is composed of four parallel metal rods with the opposing rod pairs being electrically connected. The mass fragments are separated based on their trajectories through the space between the rods. Consequently, only fragments with specific m/z ratios are detected, as they possess a stable path. By introducing changes in the electrical field, m/z range windows can be set and later analyzed. This does however impact the measurement, as the acquisition window needs to be defined before the measurements are performed [El-Aneed et al., 2009]. Alternatively, a Time of Flight (TOF) analyzer may also be used. This technology has two considerable advantages when compared to the quadrupole mass analyzer. First, it operates based on a very simple principle whereby charged mass fragments are accelerated in a long tube with a defined electric field. Secondly, this is not limited to a specific m/z range and all m/z fragments can be measured. In these instruments separation is based on the time required for the fragments to reach the detector. Therefore, smaller ions will travel faster and reach the detector before heavier ions that carry the same charge [El-Aneed et al., 2009; Jorge et al., 2016b]. The aforementioned MS techniques can be combined and then the method is referred to as tandem mass spectrometry (MS/MS). These tandem MS approaches consist of at least two m/z separation steps with a fragmentation step in between. Even though these techniques produce more complex fragmentation patterns, they can be used to investigate metabolite structures [El-Aneed et al., 2009; Jorge et al., 2016b].

Irrespective of which combination of separation, ionization and mass analyzer were used for the detection of the mass peaks. Further analysis of the resulting spectra are needed to identify and quantify the detected metabolites. These include working steps as noise filtering, baseline correction, normalization, peak picking, peak integration and peak alignment [Theodoridis et al., 2011]. In addition, spectral deconvolution can be performed to quantify and analyze metabolites from overlapping peaks [Fernie et al., 2004]. Nevertheless, the most important step is the peak identification, for which the information about the retention time, the mass spectrum and the intensity are relevant [Lisec et al., 2006; Tohge and Fernie, 2009]. In order to perform such analysis, a variety of software tools and databases, as MassBank [Horai et al., 2010], the Human Metabolome Database (HMDB) [Wishart et al., 2013] or the Golm Metabolome Database (GMD) [Hummel et al., 2013] have been developed. The detected peaks from the MS experiment can than be searched within the database using the m/z value to perform the identification of measured metabolites [Tohge and Fernie, 2009].

Nevertheless, measuring the metabolites of a sample at one or multiple successive time points does not allow the researcher to make any statements about which reactions contribute to shaping the pools of the metabolites. To overcome the limitation, isotope-labeling approaches were introduced, in which the organism of interest is fed with labeled metabolites. The most frequently used label is ^{13}C, while

^{2}H, ^{15}N and ^{18}O can be used as well [Antoniewicz, 2015; Dai and Locasale, 2017]. ^{13}C labeling experiments can be used for the estimation of carbon metabolic fluxes within an organism [Zamboni, 2011], in contrast to ^{15}N which is used to estimate nitrogen metabolism [Soong et al., 2014]. Two types of ^{13}C metabolic flux analysis (MFA) can be distinguished, as *stationary* and *non-stationary*. *Stationary* ^{13}C MFA allows to investigate alternative pathways towards a common product and are performed at isotopic steady-state. This means that the labeling pattern does not change in the course of the experiment. In contrast, *non-stationary* ^{13}C MFA allows to investigate the dynamics of the system over time with respect to the label incorporation in the metabolic pool. However, this approach requires additional computational tasks and absolute metabolite quantification [Zamboni, 2011; Antoniewicz, 2015].

3.2 The transcriptome and its measurement

The transcriptome encompasses all transcripts present in a cell including enzyme coding transcripts. The levels of enzyme-coding transcripts and their regulation may highly affect the metabolite levels within the cell [Gaudinier et al., 2015]. The production of enzymes, catalyzing metabolic reactions, starts with the transcription of the genomic sequence following three steps: initiation, elongation and termination [Wade and Struhl, 2008]. In general, the expression of a gene can be regulated through the binding of transcription factors (TF), in promoter regions or cis-regulatory elements, which can have either an activating or repressing effect [Singh, 1998; Kaufmann et al., 2010]. TFs themselves are highly regulated and can form so called gene regulatory networks, which involve feedback and feed-forward mechanisms [Macrae and Long, 2012]. During feedback regulation, the TF induces the expression of an additional gene, which will finally initiate the degradation of the TF itself [Adibi et al., 2016]. Therefore, regulation of transcripts by TFs will influence metabolite levels through the amount of transcribed mRNA and subsequently the amount of active enzymes. However, gene expression is not a completely hierarchical process. In addition, metabolites are regulating gene expression through feedback mechanisms [Gaudinier et al., 2015]. The existence of these feedback loops have been shown for the primary metabolism, especially for carbohydrates, as well as for secondary metabolites in plants [Koch, 1996; Gaudinier et al., 2015].

However, further mechanisms are responsible for fine tuning the exact amount of mRNA and subsequently the amount of enzymes. After transcription, the transcribed RNA is regulated on the post-transcriptional level through processes such as splicing or the manipulation of the stability [Floris et al., 2009]. The transcribed pre-mRNA consist of exons and introns. Through the process of splicing, introns are removed from the pre-mRNA and the mature mRNA is produced. Alternative splicing allows to produce different mRNAs from the same transcript resulting in different proteins from the same pre-mRNA. Further, the stability of mRNA is an other means whereby the amount of translated mRNA can be regulated. It was shown that the 5'-end methyl-7-guanosine cap and the 3'-poly(A) tail stabilize the mRNA, besides their function in translation [Gutiérrez et al., 1999].

The generation of functional enzymes requires the translation of the mature mRNA, which requires the presence of ribosomes and amino acid loaded tRNAs. Generally, translation is affected by the structure of the mature mRNA. The above mentioned 5'-end methyl-7-guanosine cap and the 3'-poly(A) tail promote translation, whereas hairpin loops in the secondary structure of the mRNA can block the translation process [Merchante et al., 2017]. Besides these mechanisms, there is evidence for metabolite specific translational regulation. An example is the translation of the *bZIP11* transcript, which is

repressed in the presence of sucrose. It was shown that the second of the four upstream open reading frames (uORF) is responsible for the repression. The uORF encodes for a short peptide sequence which is capable of stopping the translation within the ribosome in the presence of high sucrose concentrations [Rahmani et al., 2009]. Therefore, the signal from the transcript level propagates through the system and influences the change of metabolite levels through the enzymes in the cell.

While transcripts and their regulation shape the overall amount of enzymes within a cell, further regulatory mechanisms are responsible for the fine tuning of the metabolite fluxes through the system. These regulations are directly affecting the activity of enzymes. One regulatory mechanism is post-translational modifications (PTM), which are fast and efficient ways for the cell to modify the amount of active proteins, in contrast to performing a *de novo* synthesis of the required proteins [Friso and van Wijk, 2015]. Modifications influencing the activity of proteins are among others phosphorylation and acetylation [Friso and van Wijk, 2015; Bartel and Citovsky, 2012]. In addition, an important modification is ubiquitination which represents an efficient way to regulate the amount of protein in the cell. After the protein has been modified with an ubiquitin molecule, it is marked for degradation [Seo and Mas, 2014]. Further, the enzyme activity is regulated through product feedback inhibition or feed-forward loops. This is realized through allostery, a process by which a ligand binds at the allosteric site of the enzyme and either activates or inhibits its function [Goodey and Benkovic, 2008]. The change of activity is thereby achieved through a conformation change of the enzyme after binding the ligand [Kamata et al., 2004]. Therefore, the multitude of regulatory mechanisms necessitates the combined investigation of metabolite and transcript levels which then allows to investigate their mutual relationship and the underlying regulatory processes.

3.2.1 Transcriptomic technologies

The generation of transcript data is different from the mass spectrometry techniques described in section 3.1.2. In general, the methods can be divided into polymerase chain reaction (PCR), microarray or sequencing based. The latter two approaches allow the simultaneous measurement of up to several thousand of transcripts, and each method has its benefits and drawbacks [Malone and Oliver, 2011].

Quantitative real-time polymerase chain reaction (qRT-PCR) is based on the standard PCR concept. It requires reverse transcription of the extracted mRNA to generate cDNA. Further, primer sequences and a DNA polymerase are needed to amplify the cDNA. qRT-PCR methods can be divided into double-stranded DNA (dsDNA)-binding fluorescent dyes and fluorescent probe based approaches. The first type detects the binding of the fluorophore into the newly generated DNA and is therefore non-sequence specific [Fitzgerald and McQualter, 2014]. In contrast, fluorescent probes are oligonucleotides with a fluorophore and a quencher attached to the molecule and complementary to a region of the cDNA. The quencher absorbs the emitted light from the fluorophore. The fluorophore and the quencher are separated during the amplification which allows the detection of the free fluorophore. Both types allow the PCR reaction to be followed in real time. Further, the signal is proportional to the amount of DNA generated by the PCR [Meyers et al., 2004; Wagner, 2013]. This allows the use of the amplification curve generated during the experiment, to quantify the initial concentration of the transcript [Meyers et al., 2004]. While the method has the advantage to measure low abundance transcripts in a fast and reliable way, it is not capable of performing an analysis of the complete transcriptome.

This limitation can be overcome by the usage of DNA microarrays. While there are several different microarray platforms which differ only by design, such as printed microarrays or in situ-synthesized

oligonucleotide microarrays, the experimental procedure for all of them are very similar [Miller and Tang, 2009]. Printed microarrays can be further separated into whether cDNA probes are generated by PCR using gene specific primers or *in situ* generated short oligonucleotides [Rensink and Buell, 2005]. In both cases the probes are spotted on glass microscope slides [Miller and Tang, 2009]. In contrast, in situ-synthesized oligonucleotide microarrays are generated through directly synthesizing the probes on the chip, mostly a quartz wafer [Miller and Tang, 2009]. In general, the probe sequences are used to recognize an "unknown" sequence, called the target, where each spot on the chip contains millions of identical probe sequences [Rensink and Buell, 2005; Lowe et al., 2017]. The detection of a binding event between the probe and the target is measured by a fluorescent scanner, as all targets are labeled before hybridization. The signal strength however depends on the number of bound targets, therefore microarray platforms allow for relative quantification. Through the introduction of two different fluorescent markers, the ratio of the gene expression between experimental conditions can be directly investigated on the same chip [Rensink and Buell, 2005; Miller and Tang, 2009].

A considerable drawback of microarrays is the limited comparability and reproducibility between platforms, as different probe sequences can be spotted [Schulze and Downward, 2001; Draghici et al., 2006]. Moreover, microarray platforms are limited due to the fact that probes are designed on prior knowledge. The approach is likely to miss RNA editing events and might not detect allele-specific differences [Malone and Oliver, 2011]. Nevertheless, microarry-based gene expression analysis are still employed, as they offer a fast and efficient way to investigate the response of known genes under induced stress conditions [Pilcher et al., 2017; Yu et al., 2018]. However to overcome the above mentioned limitations, RNA-sequencing (RNA-seq) approaches can be employed. A main advantage of RNA-seq techniques is their independence of prior generated probes allowing to measure previously unknown mRNAs. Further, RNA-sequencing technologies hold the possibility of exact quantification of the mRNAs. Although, a large diversity of sequencing technologies exists, this literature review will cover only the most frequently used next-generation sequencing (NGS) technologies. Sanger sequencing was the first sequencing technology and is available since the 1970s. However, widespread use of sequencing only became available after 2005 with the introduction of the NGS method(s).

All discussed NGS methods require the synthesis of cDNA libraries through reverse transcription and the ligation of adapter sequences before sequencing [Jain, 2012; Chu and Corey, 2012; Lowe et al., 2017]. In addition, several preparation steps can be done beforehand to enhance the quality of the actual sequencing. In order to increase the sensitivity, probes binding the poly(A)-tail of mRNAs can be used to enrich the abundance of mRNAs and reduce the amount of other RNAs, such as ribosomal RNA and microRNAs. Further, RNA-seq approaches generate read-lengths shorter than the length of mRNAs, which requires their fragmentation to measure them completely. In addition, an amplification step can be used to increase the amount of low-abundant cDNAs [Lowe et al., 2017]. The resulting cDNA libraries are than provided with specific adapter sequences. For the 454-pyrosequencing technology from Roche, the pre-prepared cDNA libraries with an adapter sequence are attached to small beads. The sequences are then amplified by PCR, which should result in each bead being covered by copies of a single cDNA sequence. In each cycle, one of the four nucleotides is washed over the plate together with the two enzymes ATP sulfurylase and luciferase. When a nucleotide is incorporated into the growing DNA strand, PPi is released and converted to ATP by the ATP sulfurylase using adenosine-5-phosphosulfate. In the second step, ATP is used by the luciferase to produce luceferin which causes the emission of light and thus detected. Between each cycle, the remaining enzymes and substrates are washed off. The technology is capable of sequencing 400-500 base pair long reads in each well up

to millions of wells simultaneously [Heather and Chain, 2016; Hakeem et al., 2016].

The second frequently used sequence-by-synthesis technique is based on the principle of reversible terminator chemistry and can be found in Illumina MiSeq and HiSeq systems. After cDNA library generation, the cDNA is attached to a solid phase via the complementary adapter sequences. Clonal amplification is done by the so-called bridge amplification. The reasoning behind the name is that the DNA strands have to bend over to begin the next round of amplification. The sequencing step itself is performed through dNTPs with a fluorophore at the 3'hydroxyl position. The fluorophore serves two purposes: first it prevents the binding of further dNTPs, and secondly, the incorporated dNTP can be monitored through the emission of the fluorophore upon excitation with a laser. The fluorophore is enzymatically cleaved off before the next washing step. While the system produces shorter reads (125 to 300 base pairs depending on the actual used machine), it compensates this through the production of paired-end reads. This means that the DNA strand is sequenced from both ends, which finally improves the mapping accuracy [Heather and Chain, 2016; Hakeem et al., 2016]. Further, paired-end sequencing is more likely to estimate gene isoforms, in comparison with single-end sequencing [Salzman et al., 2011].

The methods described above represent two of the most widely used second-generation sequencing technologies. Third-generation sequencing technologies have also started to emerge. While there is no clear consensus on the separation from second to third-generation sequencing, some key points are single molecule sequencing and real-time sequencing. The biggest advantage is that no amplification step is needed, which reduces bias introduced through this step [Heather and Chain, 2016]. An example is the single molecule real time (SMRT) platform. The approach allows to monitor the extension of a DNA molecule by single dNTPs in real time. In order to do so, a single DNA polymerase is positioned at the bottom of a well, over which fluorophore labeled dNTPs are washed. The incorporation of dNTPs into the growing DNA is detected by a laser. The laser passes through an aperture of a diameter smaller than its wavelength causing its light intensity to decay exponentially. This allows to monitor the incorporation of a single dNTP into the DNA at the bottom of the well without the interference from other labeled dNTPs .[Heather and Chain, 2016].

Independently of the used platform, further computational tasks have to be performed before the data can be evaluated in their biological context. These are quality control of the read and the alignment of the read to a reference genome. Careful consideration is required before selecting a method, as the resulting read length need to be evaluated with respect to the organism studied [Jain, 2012; Chu and Corey, 2012; Conesa et al., 2016; Lowe et al., 2017]. Overall, methods for the investigation of transcripts have enhanced the general knowledge of biological processes and allow for the comparison to other data types. The investigation of gene expression in concordance with metabolomics data is of importance to understand underlying biological processes and to investigate regulatory processes between the two layers. All of the microarray and sequencing approaches provide the opportunity to investigate gene expression in detail and on a genome-scale level.

3.3 Data reduction and regression approaches for transcriptomic and metabolomic data analysis

In section 3.1.1 stoichiometric and kinetic modeling were introduced allowing for the simulation of metabolite fluxes and metabolite concentrations, respectively. In contrast, statistical modeling can be

used to investigate large-scale data sets, as those originating from high-throughput measurements. It allows to analyze the relationship between sets of variables and to test how well the models describe the relationship, by estimating confidence intervals and/or p-values. Therefore, statistical modeling can be used to analyze transcriptomic and metabolomic data and to investigate the mutual regulatory mechanisms of transcripts and metabolites. However, the combined analysis poses a challenge. These data sets often have high dimensionality and differ in the number of measured components since several thousand genes can be detected [Meyers et al., 2004; Jain, 2012] compared to around a thousand metabolites [Giavalisco et al., 2008]. Generally, any multivariate statistical method used to investigate genes and metabolites aims to reveal the relationship and association between the two layers. In order to do so, the approaches can be classified depending on whether they are designed to investigate the relationship of a single gene to a single metabolite, multiple genes to a single metabolite, a single gene to multiple metabolites or multiple genes and multiple metabolites.

The investigation of a single gene and a single metabolite can be performed through pairwise Pearson or Spearman correlation [Tohge et al., 2015; Cavill et al., 2016]. However, the high number of correlations between transcripts and metabolites renders the analysis challenging to directly determine cellular mechanisms from the investigation [Urbanczyk-Wochniak et al., 2003]. Further, it was shown that the direction and magnitude of transcript-metabolite correlation changes between experimental conditions requiring additional caution when interpreting the results [Bradley et al., 2008]. Nevertheless, this type of analysis helps to elucidate the relationships between genes and metabolites with additional experimental designs, particularly for gene function annotation. In order to do so, the combination of metabolite and transcript levels are compared between different genotypes (e.g. knockout mutants or natural variants) associating genotypical and phenotypical changes [Tohge and Fernie, 2010, 2012]. However, the information gain by one-to-one comparisons is limited in system-wide investigation, as the multitude of regulation levels can not be fully elucidated in the context of single metabolite and transcript associations.

This limitation can be overcome by incorporating more information to gain a system-wide overview in order to reveal underling regulatory mechanisms. To do so, a more elaborate investigation involves the study of associations between multiple genes and a single metabolite or between multiple metabolites and single gene. However, this requires additional statistical approaches to reduce the dimension of one data set. This can be performed with principle component analysis (PCA). PCA uses an orthogonal linear transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called PCs. The PCs are ordered according to the variance they explain [Wold et al., 1987]. In the study of Inouye et al. [2010], the authors identified modules of co-expressed genes on which they perform PCA. The association (calculated with Spearman correlation) between the first principal component and metabolic data profiles is than used as a means to determine which metabolites are influenced by the genes of the module [Inouye et al., 2010].

In addition to combining PCA and correlation approaches to investigate associated transcripts and metabolites, it is possible to use regression based approaches. The standard regression model can be written as $Y = X\beta + \varepsilon$, where Y is the response vector, X is a matrix of values of predictive variables, β contains the parameters and ϵ is a noise vector [Bickel et al., 2009]. In concept, transcript levels can be used to predict metabolite levels and compare them to measurements to verify the prediction. This allows to estimate which transcripts influence metabolism [Auslander et al., 2016]. In addition, it has been shown that metabolite levels can be used with the same approach to investigate regulation from metabolites on the transcript level [Kochanowski et al., 2017]. The regression approach tries to

find parameters such that Y can be explained by $X\beta$. This can be achieved by finding the maximum-likelihood estimator of β which can be estimated as follows: $\hat{\beta} = (X^T X)^{-1} X^T Y$. However, this requires that $X^T X$ is invertible which is not the case when the number of variables (p) is larger than the number of observations (n), resulting in a $n < p$ problem. Therefore, regularization or dimension reduction approaches have to be used to learn regression models [Johnstone and Titterington, 2009]. In Auslander et al. [2016] this has been achieved by selecting a pair of genes for each metabolite, which showed high positive or negative correlation to the metabolite, before predicting each metabolite level separately. In contrast, Kochanowski et al. [2017] performed a linear regression without regularization to estimate metabolites affecting transcriptional regulation. The approach relates one or two metabolites to one promoter restricting the analysis to a small set of metabolites and promoters of the central metabolism. Discussed examples were allowed estimation of the effect of small sets of metabolites or transcripts onto the other level. However, the knowledge gain is still limited to a smaller number of regulatory mechanisms and still lacks a system-wide estimation.

Therefore, to answer questions of system-wide regulation one needs to investigate the relation of multiple genes to multiple metabolites. A possible strategy is to use Partial Least Squares (PLS) regression and its extension OPLS and O2PLS [Bylesjö et al., 2007]. PLS aims to find the relationship between two matrices X and Y by estimating the direction in X that explains most of the variance in Y [Boulesteix and Strimmer, 2007]. Due to its multivariate nature, PLS regression is difficult to interpret. The orthogonal projections to latent structures (OPLS) was designed to improve the interpretation of the regression. The approach removes the variation from X which is uncorrelated (orthogonal) to Y. The advantage compared to PLS is twofold. First the orthogonal part of X can be separately investigated. Secondly, and more importantly the removal of uncorrelated variation increases the interpretation. Subsequently, O2PLS is an extension of OPLS. The O2PLS model for a combined analysis of transcript and metabolite data consist of three different parts. The first part is the joint part between the matrices X and Y. This part can be seen as the integration of both data sets and contains the information from X that explain Y and the other way around. The second part is the orthogonal part containing the underlying latent variables responsible for the unique systematic variation in X and Y, respectively. The last part captures the noise in X and Y. The joint part is of special interest in this approach, as it allows to predict metabolite profiles from transcript data and the other way around [Trygg and Wold, 2002; el Bouhaddani et al., 2016; Bylesjö et al., 2007]. The O2PLS approach has been shown to be applicable to the large-scale metabolomic and transcriptomic data introduced in Inouye et al. [2010], reproducing the results without the need to perform PCA beforehand.

Beside regression based approaches, canonical correlation analysis (CCA) can be used to perform a joint investigation of transcript and metabolite data. Given two data sets X and Y, CCA finds the canonical variates, $U = a'X$ and $V = b'Y$, so that the correlation between U and V is maximized [Hotelling, 1936]. The advantage of CCA is that it is invariant with respect to transformation of the variables. A clear drawback of CCA is that it requires the calculation of the inverse of XX^T. Therefore, CCA is not applicable with data sets in which the number of variables is larger than the number of observations ($n < p$ problem), similar to regression approaches. Again, dimension reduction has to be performed to apply CCA to transcriptomic or metabolomic data. A possible workflow has been shown in Jozefczuk et al. [2010] in which each experimental condition was separately investigated. While this does not allow to investigate metabolite and transcript associations over multiple conditions which would result in an understanding of general regulatory mechanism, it distinguishes between condition specific effects. Further, a downstream analysis can retrieve similarities and distinct differences

between experimental conditions. An additional approach for the investigation of multiple genes and metabolites has been illustrated by Oliveira et al. [2015]. The authors used Bayesian inference to estimate the position of the metabolite in relation to the regulatory protein target of rapamycin complex 1 (TORC1) which could be downstream, upstream, parallel or unrelated to TORC1. Through the usage of time-series measurements of metabolites and transcripts and the integration of additional network information, the approach allowed to identify metabolites affecting TORC1 regulation or being affected by it. The integration of prior knowledge with transcriptomic and metabolomic data in a Bayesian inference approach allows to gain specific insights into regulatory mechanisms. However, the approach limits the investigation towards the relationships associated with the integrated knowledge.

The presented approaches give an illustration of the diverse available statistical methods to probe the relationship between metabolites and transcripts. While the investigations of pairwise relationships between genes and metabolites can be helpful to understand specific parts of the metabolic network, it does not allow to make statements about system-wide changes. However, this is of importance when investigating the reaction to environmental changes on a system-wide level. Therefore, the analysis of multiple genes to single metabolites or multiple metabolites to single genes enables to examine regulatory mechanisms between different layers. This increases the knowledge towards a systemic understanding of the association of transcriptional levels and metabolism. Finally, investigations of multiple genes to multiple metabolites can further elucidate system-wide regulatory mechanisms.

3.4 Thesis outline

The main aim of my thesis was to further explore the existing methods as well as to implement novel methods for the evaluation and integration of metabolomic and transcriptomic data while focusing on regulatory mechanism. The three result chapters contain published studies in the field of systems biology. In chapter 4, I investigated a previously published method from observability theory. The implementation of the approach was published in the study of Liu et al. [2013] and is capable of identifying so called sensors. These sensors represent nodes of the network of interest that are sufficient to reconstruct the internal state of the system. While the study indicated that the approach can be useful for the investigation of metabolomics studies, it did not provided evidence for this claim. However, I was able to show that identified sensor metabolites are highly correlated with each other in comparison to non-sensor metabolites. Therefore, metabolite levels reflect the role of components in the observability of the system. This approach might help biologists to focus on a specific set of metabolites to describe and compare the phenotypic state of an organism [Schwahn et al., 2016].

The remaining two studies introduced novel approaches, capable of investigating underlying regulatory mechanisms, either between metabolic reactions or on the level of transcription and post-transcription. There is increasing evidence that further regulation at the level of metabolic reactions exists, as transcript levels are not sufficient to explain the observed metabolic changes [Daran-Lapujade et al., 2007; Chubukov et al., 2013], indicating there must be further regulation at the level of reactions through reaction coupling. While reaction coupling can be investigated in genome-scale metabolic networks, the networks may be incomplete and therefore might not reflect the actual complexity of the regulation [Becker et al., 2006; Burgard et al., 2004; Millard et al., 2017]. In chapter 5, I proposed a new method, which estimates how many reaction couplings take place within an investigated system. The novelty here is that no metabolic network representation is needed and that the

estimation is purely based on measured metabolite levels. The method is capable of comparing the degree of regulation at the level of reactions between two or more organisms. I was able to detect a loss of regulation that occurred during the cultivation of wheat and tomato [Schwahn et al., 2017a].

In chapter 6, I proposed two novel methods to integrate metabolomic and transcriptomic data. The methods are able to group pairs of metabolites into two categories, either as being mainly regulated at the transcriptional level or on the post-transcriptional level. The approach can be used to integrate large-scale data sets from high-throughput experiments while performing a dimension reduction only on the transcriptomic data. I showed that the categorization of metabolite pairs being associated due to transcriptional or post-transcriptional regulation is in agreement with previous published results, as well as known regulatory mechanisms. Overall, these studies and the proposed methods provide the means for a deeper and more careful investigation of metabolomic profiles. This can increase our knowledge twofold: firstly, by identifying phenotypically relevant metabolites and secondly, by detecting regulatory effects between metabolites. Finally, the studies are summarized and future developments are proposed in chapter 7.

4 Observability of plant metabolic networks is reflected in the correlation of metabolic profiles

Publication: Plant Physiol. 2016 Oct;172(2):1324-1333. Epub 2016 Aug 26.

Authors: Kevin Schwahn[1,2], Anika Küken[1], Daniel J. Kliebenstein[3,4], Alisdair R. Fernie[2], Zoran Nikoloski[1]

Affiliations: [1]Systems Biology and Mathematical Modeling Group,
Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, Potsdam-Golm, Germany

[2]Central Metabolism Group, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany

[3]Department of Plant Sciences, University of California, Davis,
One Shields Avenue, Davis, CA 95616

[4]DynaMo Center of Excellence, University of Copenhagen, Thorvaldsensvej 40, DK-1871, Frederiksberg C, Denmark

Contact: *nikoloski@mpimp-golm.mpg.de

Abstract

Understanding whether the functionality of a biological system can be characterized by measuring few selected components is key to targeted phenotyping techniques in systems biology. Methods from observability theory have proven useful in identifying sensor components that have to be measured to obtain information about the entire system. Yet, the extent to which the data profiles reflect the role of components in the observability of the system remains unexplored. Here we first identify the sensor metabolites in the model plant *Arabidopsis thaliana* by employing state-of-the-art genome-scale metabolic networks. By using metabolic data profiles from a set of seven environmental perturbations as well as from natural variability, we demonstrate that the data profiles of sensor metabolites are more correlated than those of non-sensor metabolites. This pattern was confirmed with *in silico* generated metabolic profiles from a medium-size kinetic model of plant central carbon metabolism. Altogether, due to the small number of identified sensors, our study implies that targeted metabolite analyses may provide the vast majority of relevant information about plant metabolic systems.

4.1 Introduction

Systems biology aims at developing models that allow for a complete characterization of how the inputs and outputs of a biological system are interconnected and jointly relate to the molecular phenotypes. The experimental systems biology studies attempt to obtain a substantial coverage of the (molecular) components of a biological system using various technological platforms, such as: transcriptomics [Weber et al., 2007] metabolomics [Fiehn, 2002], and, more recently, phenomics [Araus and Cairns, 2014]. The aim of these research efforts is to utilize the read-outs about the components for estimating how the biological system functions.

However, while these efforts are rapidly becoming faster and cheaper, they still encounter both financial and logistical problems when attempting to scale up to measure large populations or the vast space of conceivable physiological environments. These problems quickly become irresolvable for any studies attempting to combine genetic and environmental variation in the same system. Thus, until the technical problems are removed, alternative solutions are in demand that can allow as much of the system to be measured (i.e. observed) as possible. Therefore, we are faced with the question: Is it possible to identify a subset of transcripts or metabolites that can provide complete information about an investigated system?

One way to identify these subsets is based on networks structures generated by systems biology approaches. This line of research aims at finding a small number of molecular components (with respect to what can be measured) whose measurement can characterize the internal state of a biological system. Given the myriad of output components from any biological system (e.g. generated within a plant leaf cell and exported to any other tissue type), it is of great interest to determine the number and the identity of these output components that may provide insights in the state of the system. However, components deemed as outputs of a modeled biological system are usually not external to the system, but rather actively participate in shaping the levels of its underlining components. For instance, amino acids are used to build proteins which, in turn, drive the entirety of metabolism, including amino acid and sugar metabolism that provide the energy and building blocks to create the plant cell wall [Cosgrove, 2005; Singh and Ghosh, 2006]. Therefore, the connectivity of components due to regulatory,

signaling, and metabolic interactions must be considered when determining the sensor components.

Metabolic networks are among the best described networks in systems biology to test our ability to identify metabolites that can serve as sensors to describe metabolism. We would like to emphasize that the concept of sensor metabolites does not correspond to the *in vivo* notion of sensing and signaling metabolites. Sensing and signaling metabolites are involved in coregulating and integrating the metabolic status with other cellular events [Templeton and Moorhead, 2004]. Our concept of sensor metabolites is that the metabolites would need to be measured by the researcher to acquire the majority of information present in the sample.

A metabolic network of a given cellular system consists of the entirety of biochemical reactions interconverting nutrients obtained from the environment into basic and more complex building blocks used to create the cell and allow it to defend itself. The components of a metabolic network are, therefore, the metabolites and the accompanying conversion reactions. These components are fully specified by the levels of all the metabolites and the rates/fluxes of all reactions. While the levels of many metabolites can be determined with modern metabolomics technologies [Goodacre et al., 2004], the reaction rates cannot be measured but are estimated from the combination of labeling and modeling [Kauffman et al., 2003; Nöh et al., 2007]. Recent advances in modeling of plants have resulted in genome-scale metabolic networks for a variety of species, from *Arabidopsis thaliana*, as a plant model, to maize and rice, as important agronomic crops [de Oliveira Dal'Molin et al., 2010a,b; Saha et al., 2011; Seaver et al., 2014].

Well-established methods from control theory utilize network structure to determine the sensors that must be measured to observe the internal state of a system, biological or otherwise [Liu et al., 2011, 2013; Jha and van Schuppen, 2001; Rios et al., 2013]. These methods are not concerned with calculating the internal states from the sensors, but determining if the system is observable with particular components. For nonlinear biological systems, such as metabolism, obtaining the internal state from the sensors is still a challenging problem [Chaves and Sontag, 2002]. The issue of determining the set of metabolites that needs to be measured in a labeling experiment to characterize a unique flux distribution of a given system has been recently tackled in the framework of constrained-based modeling [Chang et al., 2008].

Here, we address the observability problem from a data-driven perspective: To begin, we apply the graphical approach of Liu et al. [2013] to large-scale plant metabolic networks. We then investigate if, and to what extent, the data profiles about metabolites predicted as sensors relate to the rest of the metabolites in the network. In this way, we aim to bridge the gap between the existing powerful control-theoretic methods and the plethora of accumulated data from metabolomics studies. To inspect the model-based effects in the identification of sensor metabolites, we tested the robustness of the findings with two different models that guarantee good coverage with the metabolomics data. In addition, we used a medium-scale kinetic model for central carbon metabolism to further strengthen our findings from the large-scale models. The findings are further discussed with respect to the role of sensor metabolites as dead-end metabolites in the respective metabolic networks (with and without consideration of biomass reactions, used in the simulating growth). The small number of identified sensor metabolites in relation to the size of the entire metabolic network suggests that targeted metabolite analyses could provide the vast majority of relevant information about plant metabolic systems and could prove effective in strategies for crop improvement [Gu et al., 2010, 2012; Lu et al., 2011; Fernie and Schauer, 2009].

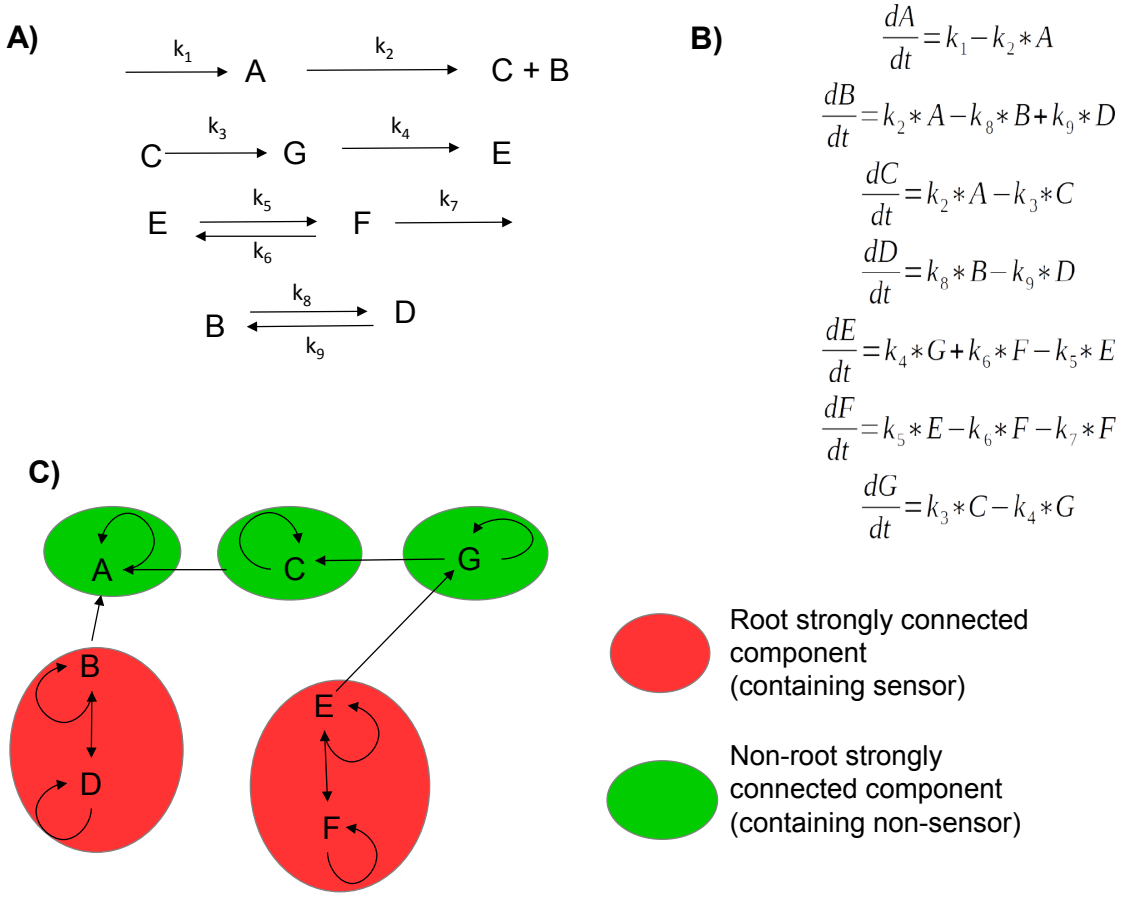

Figure 4.1: **Schematic overview of the implemented algorithm, adapted from Liu et al. [2013].** **A)** Example of a metabolic network with nine irreversible reactions. **B)** System of differential equations for the change in concentration for each metabolite (A to D) in the network shown in (A) assuming mass action kinetic. **C)** Inference graph for the metabolic network and system of differential equations in A and B. Node u is connected by a directed edge to node v if metabolite v occurs in the differential equation for metabolite u from B. The green circles represent non-root SCC, whereas the red circles indicate root SCC. Each node in a root SCC can act as a sensor node.

4.2 Materials and Methods

Our analysis is based on the graphical approach of Liu et al. [2013]. The sensor metabolites can be determined by building the inference graph obtained from a given network of biochemical reactions under the assumption that their rates are described by mass action kinetics. The nodes in the inference graph are given by the metabolites. For instance, the network on Figure 4.1A contains seven metabolites, denoted by A–G, transformed via nine reactions with rate constants $k_1 - k_9$. A node (i.e. metabolite) u is connected by a directed edge to node v if metabolite v occurs in the differential equation for metabolite u. To illustrate the building of the inference graph, we again turn to the network of biochemical reactions in Figure 4.1: Because A appears on the right-hand side of the differential equation for A (i.e. $\frac{dA}{dt}$ on Figure 4.1B), there is a directed edge from A to itself (Figure 4.1C). Similarly, there is a directed edge from node A to node B because A appears in the differential equation for metabolite B. The inference graph can be decomposed into its strongly connected components (SCC). A SCC is the maximal subgraph for which there are directed paths from every node to all others. For instance, nodes B and D form a SCC because there is an edge from B to D as well as from D to B. However, E and C are not in a SCC because there is no directed path from C to E, although there is a path from E to C (Figure 4.1C). If a SCC does not have an incoming edge, it is referred to as a "root" SCC. In our toy example, B and D as well as E and F form two root SCCs, while A, C, and G form three non-root SCCs.

Liu et al. [2013] showed that the sensors are located in this set of nodes in the root SCCs. The set of nodes obtaining by selecting at least one node from each root SCC then allows complete observability of the system. A similar framework has been also applied and discussed in Rios et al. [2013]. The approach can be readily applied to any genome-scale metabolic network because the inference graph can be built only from the stoichiometric matrix, as input. To determine the edges which start at a node u, it suffices to identify the substrate metabolites of the reactions in which the metabolite u participates as a substrate or product. The substrates of a reaction are readily given by the negative entries of the corresponding reaction vector in the stoichiometric matrix. For instance, node B participates in reactions with B, D, and A as a substrate, and, thus, there are directed edges to these nodes from B. We used the R package igraph [Csardi and Nepusz, 2006] to build the inference graph and to find its (root) SCCs.

A root SCC may not consists of a single metabolite, as is the case on the toy network in Figure 4.1C. In this case, for the root SCC consisting of B and D, any of the two can serve as a sensor. We applied the graphical approach to two genome-scale metabolic networks of *A. thaliana*, the bottom-up assembled Arabidopsis core model, AraCORE [Arnold and Nikoloski, 2014], and the Arabidopsis model from PlantSEED [Seaver et al., 2014], referred to as AraSEED. Both networks cover pathways of plant primary metabolism. We analyzed these models, whose characteristics appear in Supplemental Table 8.1.7, with and without consideration of biomass and sink reactions. The sensor metabolites were selected from the root SCCs as those that could be mapped to the available metabolic profiles. In our study, this resulted in a single sensor node identified per root SCC (see Supplemental Tables 8.1.1 and 8.1.3 for lists of identified sensors in the two models and Supplemental Tables 8.1.5 and 8.1.6 for lists of mapped metabolites).

To relate the predicted sensors to metabolic measurements, we obtained metabolic profile data from Caldana et al. [2011] generated by gas chromatography-mass spectroscopy (GC-MS). This metabolic data set consists of 91 metabolites measured under the following conditions: 21° C at 75 $\mu E\, m^{-2} s e c^{-1}$,

150 $\mu E\, m^{-2}\, s\, e\, c^{-1}$ light intensity and darkness, 4° C at 85 $\mu E\, m^{-2}\, s\, e\, c^{-1}$ light intensity and darkness, 32° C at 150 $\mu E\, m^{-2}\, s\, e\, c^{-1}$ and darkness. Therefore, the analyzed data set consisted of metabolic time series covering 20 time points and gathered under seven conditions. In addition, to augment the set of tested conditions, we used metabolic data profiles from a study of natural variation in central carbon metabolism of *A. thaliana* [Sulpice et al., 2013]. In this study, the data profiles of 45 metabolites were measured in 97 *A. thaliana* lines in three conditions, namely 8h of light with high nitrogen supply, 12h of light with high nitrogen supply and 12h of light with low nitrogen supply. Metabolite data were acquired using GC-MS technology. A detailed description of the plant growth conditions and experimental design can be found in the "Materials and Method" section of Sulpice et al. [2013]. The two studies whose data sets we used here performed their GC-MS experiments as outlined in Lisec et al. [2006].

These data sets allow first insights, to our knowledge, into how the sensors relate to the rest of the measured metabolome under a variety of genotypes, environmental conditions, and over time. For the statistical analysis, we tested for differences in the means of correlation values between the two groups of sensor and non-sensor metabolites by two-sided t-test at significance level of $\alpha = 0.05$. A graphical representation of the complete workflow applied in this study is visualized in Supplemental Figure 8.1.1.

Genome-scale metabolic networks are open systems, in contrast to the closed systems (i.e. without in- and out-flux reactions) considered by Liu et al. [2013] and Rios et al. [2013]. Because an open system has additional self-edges at the output metabolites in the inference graph. These have no effect on the identification of root SCCs (see Supplementary information Liu et al. [2013]). Figure 4.1 illustrates an open system in which the root SCCs remain unaffected if the import and export reactions are removed. To identify dead-end metabolites given a large-scale network, we used the COBRA toolbox function *removeDeadEnds* in MATLAB [Schellenberger et al., 2011].

4.3 Results and Discussion

4.3.1 Number and position of sensor metabolites in models of plant primary metabolism

By applying the graphical approach to identify root SCCs in the Arabidopsis core model (AraCORE, [Arnold and Nikoloski, 2014]), we found 23 sensor metabolites listed in Supplemental Table 8.1.1. Aside from two sugars, trehalose and cellulose, and nucleoside triphosphates, the remaining sensor metabolites were amino acids. The metabolomics data set of Caldana et al. [2011] contained the metabolic profiles of 11 of the identified sensor metabolites. Overall, 30 of the 91 measured metabolites could be mapped to AraCORE, as indicated in Supplemental Table 8.1.5 that includes the metabolites used as sensors and non-sensors for the investigation of this model. Consideration of biomass and sink reactions in the model led to the identification of only 15 sensor metabolites, consisting of the amino acids and cellulose (see Supplemental Table 8.1.2 – "AraCORE Sensors with Biomass Function"). The finding that the sensors identified upon consideration of biomass also act as sensors when biomass is excluded was in line with the observation that the biomass reaction includes all amino acids, alongside cellulose and nucleotides.

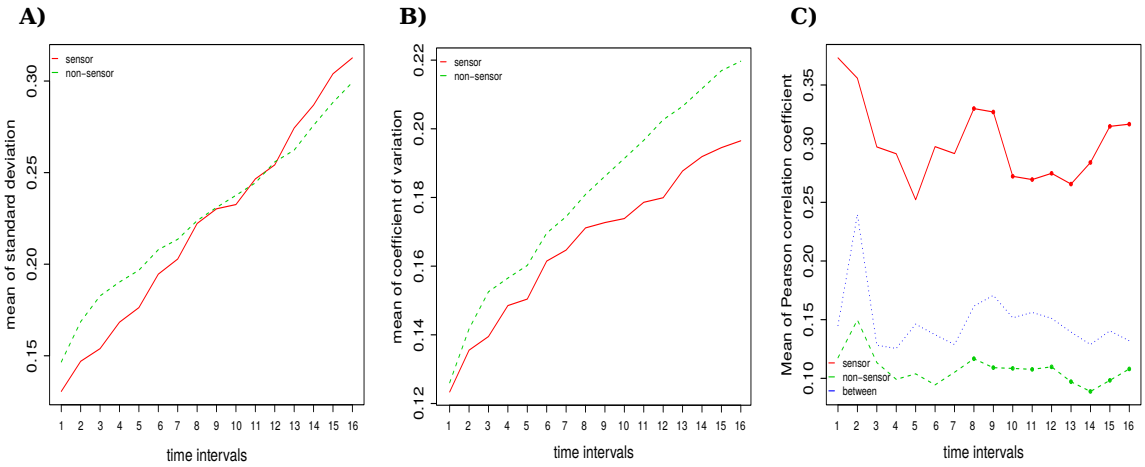

Figure 4.2: **Statistical comparison of sensors and non-sensors in the AraCORE model.**
The x axis represents the investigated time interval, from 1 to 16. The y axis represents values for the three statistics, respectively: **A)** SD ; **B)** CV; and **C)** Pearson correlation of sensors and non-sensors. The red line corresponds to the values for the statistics between sensor metabolites, while the green line corresponds to values between non-sensor metabolites. The blue line in C is used for the correlation between sensors and non-sensors. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and non-sensors.

Additionally, we also considered the Arabidopsis model downloaded from PlantSEED (AraSEED [Seaver et al., 2014]). We identified 198 sensor metabolites, given in Supplemental Table 8.1.3, of which 10 could be mapped to the metabolomics data. Overall, we were able to map 47 metabolites to the measured data. The list of all metabolites identified as sensor and non-sensor in the investigation of the AraSEED model is provided in Supplemental Table 8.1.6. In agreement with the AraCORE model, the sensors again included amino acids and sugars, in addition to variety of complexes with Coenzyme A and Plastoquinone. In brief, the findings from the two models were similar in that all models have root SCCs that largely overlap sugar and amino acid metabolism. However, the small number of mapped metabolites in comparison to the size of the models employed is a challenge, largely due to the limitations of the current metabolomics technologies. For instance, a quarter of the detected analytes could not be annotated to known metabolites; moreover, secondary metabolites could not be mapped in all models, because some of the models used in this study include pathways of central metabolism.

4.3.2 Data profiles of sensor metabolites show stronger correlations than non-sensor metabolites

The underlying approach states that information from all root SCCs allows the reconstruction of the state of the system. A minimum set of sensor metabolites can then be used to specify the metabolic profiles of the sensors and the rest of the network. It is important to emphasize that the metabolites from the root SCCs, containing the sensors, can be connected to different non-root SCCs. Therefore, one may expect that there is a relation within sensors based on whether they are connected to the same non-root SCCs (see Supplemental Figure 8.1.1 for illustration). If all nodes in a non-root SCC have a directed path to sensors in two SCCs, a single sensor may suffice to reconstruct the state of the non-root SCCs; in this case, the other sensor will be needed to describe its own profile. For the investigated networks, the majority of the identified sensor metabolites were connected, via a directed path, to

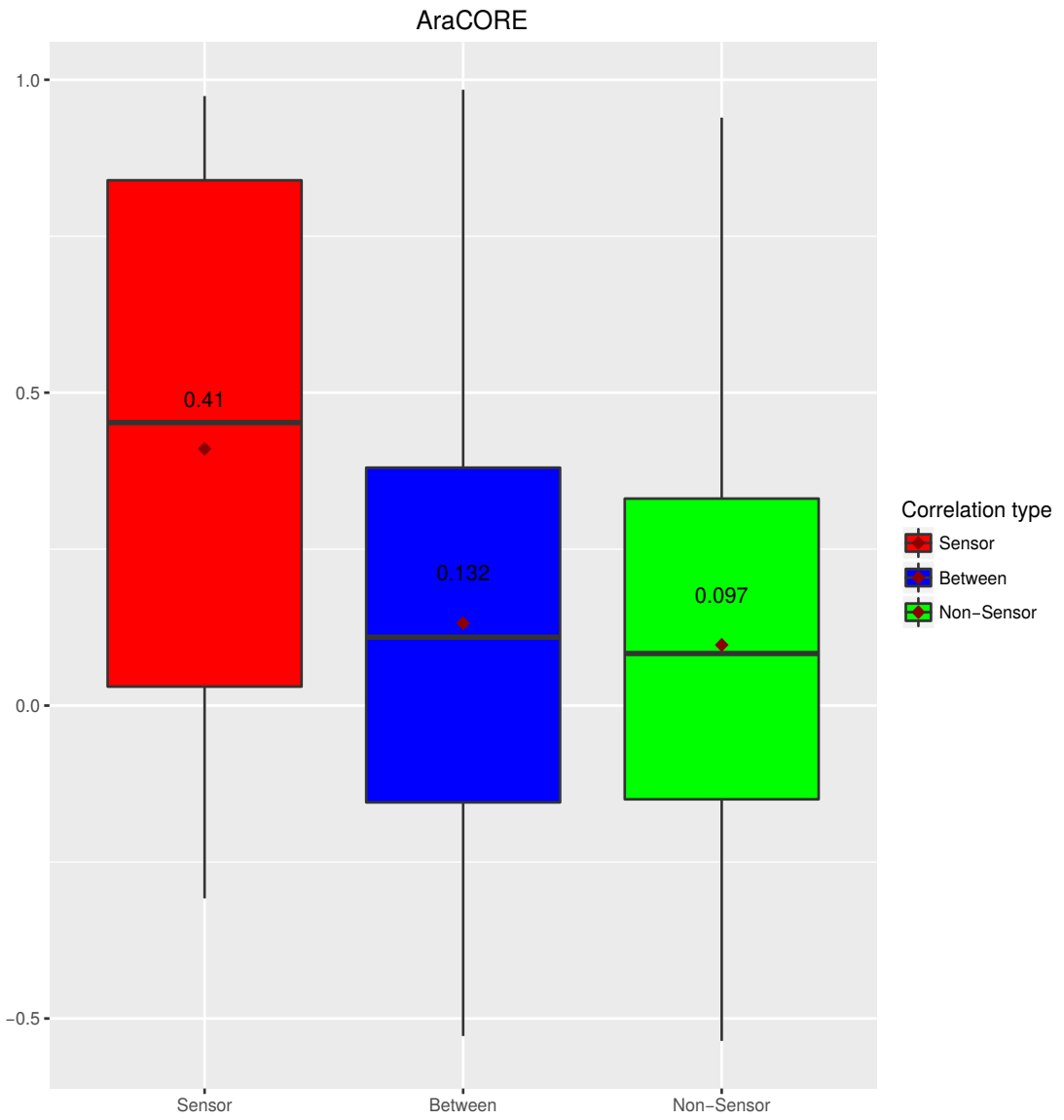

Figure 4.3: **Distribution of correlation values of the AraCORE model.**
Box plots of the distribution of correlation values between sensors, between sensors and non-sensors, and between non-sensors are colored in red, blue, and green, respectively. The mean value is given above the square symbol, while the median is given by the solid line.

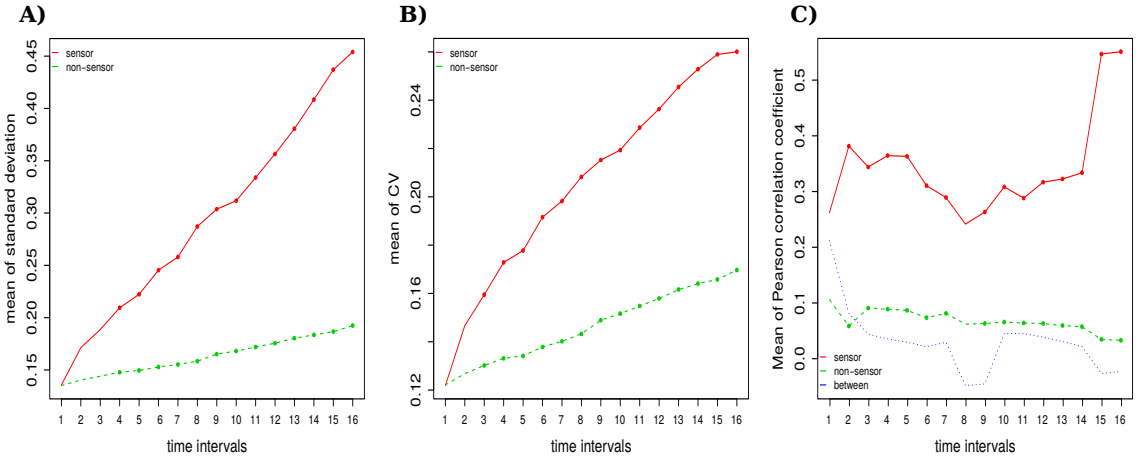

Figure 4.4: **Statistical comparison of sensors and non-sensors in the AraSEED model.**
The x axis represents the investigated time interval, from 1 to 16. The y axis represents values for the three statistics, respectively: **A)** SD ; **B)** CV; and **C)** Pearson correlation of sensors and non-sensors. The red line corresponds to the values for the statistics between sensor metabolites, while the green line corresponds to values between non-sensor metabolites. The blue line in C is used for the correlation between sensors and non-sensors. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and non-sensors.

the same non-root SCC. Therefore, sensors in different root SCCs detect the same network and each could be employed to reconstruct the state of the non-root SCC. Therefore, if the data profiles of a sensor metabolite can be used to reconstruct the profiles of the non-sensor metabolites, it may be expected that sensor metabolites are more correlated to each other than to the rest of the metabolites; by corollary, for non-sensor metabolites, it may be expected that they are less correlated to each other than to the sensor metabolites. Within the AraCORE model all sensors are connected to the same non-root SCC, whereas in the AraSEED model 35 of the 198 sensors were not connected to the largest non-root SCC. Out of the 35 sensors only glucose and fructose were mapped to the data. We did not observe a different behavior with respect to the findings from the previous analysis for two types of sensor groups (see Supplemental Figure 8.1.4).

To empirically test these sensor hypotheses and their biological utility we used time-series metabolomics data from *A. thaliana* Col-0 exposed to seven different environments. To this end, we determined the correlation for each pair of measured metabolites over all conditions; we then divided the resulting correlation values in three categories: (1) between two sensors, (2) two non-sensors, and (3) between a sensor and a non-sensor metabolite. Since the available time-series data captured the response to the applied perturbations caused by the different light and temperature conditions across different time scales, we determined the correlation between the time series with consideration of different time points (i.e. intervals). More specifically, we determined the correlations by using k (with $5 \leq k \leq 20$) consecutive time points from the experimental measurements, starting with the first time point (Figure 4.2). This results in 16 time intervals, so that the first interval consists of the first five measured time points and the last of all 20 time points. In addition, we investigated the correlation obtained by jointly considering the data from all time points and conditions.

This analysis provided the distributions of correlation values across the three classes of metabolite pairs in a given time interval over all considered conditions. We then tested the null hypothesis that the means of the distributions do not statistically differ between the classes of metabolite pairs, by

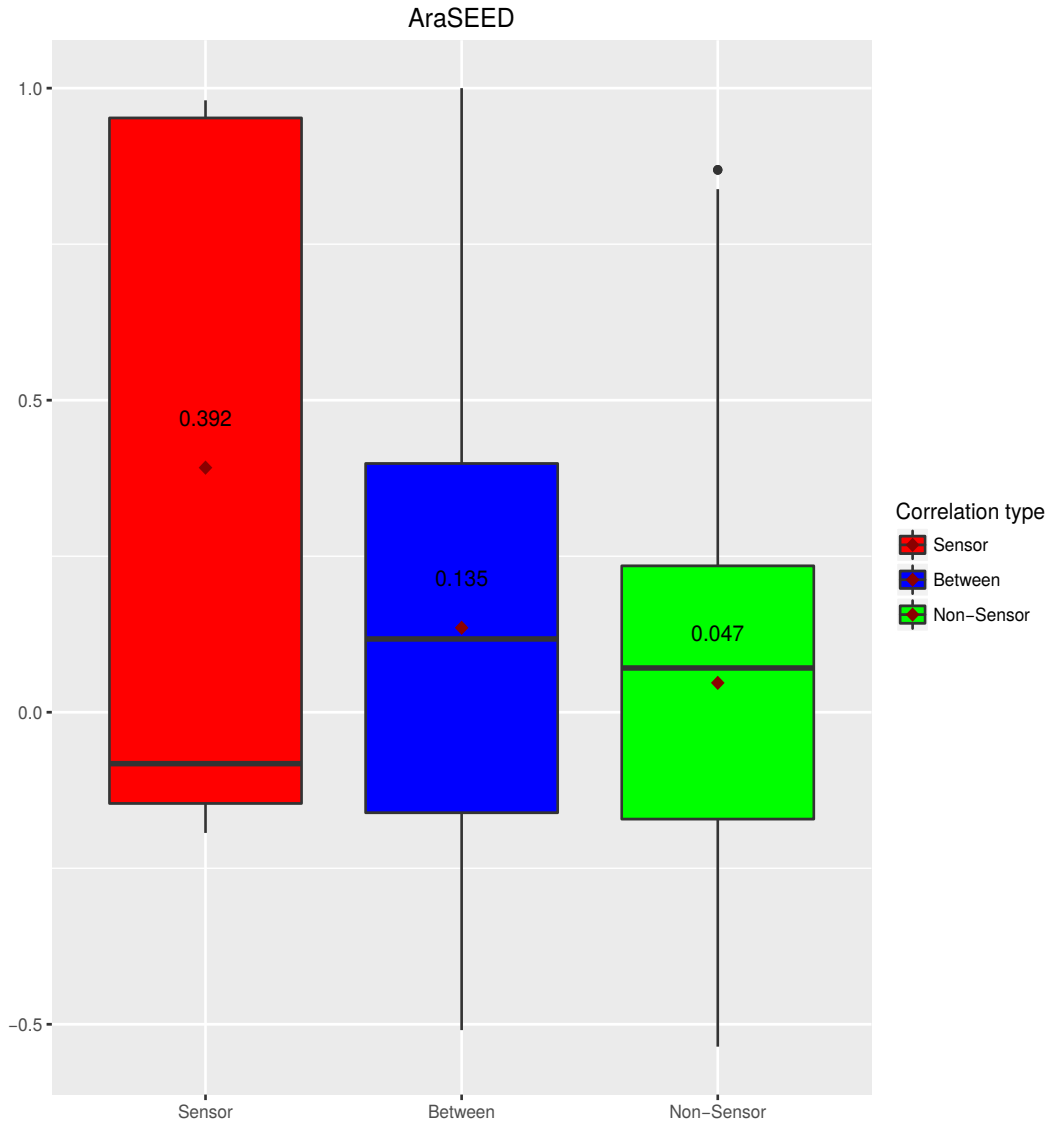

Figure 4.5: **Distribution of correlation values of the AraSEED model.**
Box plots of the distribution of correlation values between sensors, between sensors and non-sensors, and between non-sensors are colored in red, blue, and green, respectively. The mean value is given above the square symbol, while the median is given by the solid line.

applying two-sided *t*-test. In accordance with the observation that the majority of the identified sensor metabolites were connected to the same (non-root) SCCs, for the AraCORE model, we found that the mean of correlations between sensor metabolites was greater than the mean of correlations between non-sensor metabolites in 9 of the 16 investigated intervals. The statistical significance in the later time points is due to the larger power of the test due the larger number of data points available [Schönbrodt and Perugini, 2013]. We also observed that the mean of correlations between sensor and non-sensor metabolites was greater than the correlations between non-sensor metabolites, but smaller than the correlations between sensor metabolites only. These results were reproducible if all time points and conditions were jointly used (Figure 4.3).

However, another possible source of this result is that the metabolic profiles of the sensor metabolites have lower variability than non-sensors, and, thus, show higher correlations. To test this hypothesis, we first determined the distributions of standard deviation (SD) and coefficient of variation (CV), for the sensor and non-sensor metabolites equivalent to the set-up for investigating the correlation values. We then tested if the means of each measure of variability differed between the classes of metabolites. The means of the CV and the SD were not statistically different between the non-sensors and sensors; thus, differences in the variation of sensors and non-sensors were likely not causing the difference in the correlation structure. Therefore, we concluded that the observed difference in correlation was a result of the position of the sensor nodes in the network and not due to smaller variability.

For a comparison of genome-scale models, we investigated the relationship of sensors and non-sensors in the AraSEED model. The mean correlations of the sensors were significantly different and larger than those of the non-sensors in 13 of 16 time intervals (see Figure 4.4), thus conforming our previous findings. This was additionally confirmed through the investigation of all time points and conditions (Figure 4.5). The correlations in sensors were significantly higher than in non-sensors. However, the results of the SD and the CV were in contrast to our previous results: In the majority of time intervals, we found significantly higher values in the sensors, than for the non-sensors. This is likely due to the difference in the number of sensors and non-sensors mapped from the metabolomics data in the two models.

Altogether, we demonstrated that, with the used data set, sensors show larger correlation than between non-sensors and sensors, and that that the latter is greater than the correlation within non-sensors. We also showed that these findings remained largely unaltered when models of different size and structure are explored. The evidence indicates that in the case of AraCORE, these finding is likely not related to the variability in the metabolic profiles. In addition, we investigated correlation of the metabolic traits gathered in a study by Sulpice et al. [2013]. The data were obtained under three different growth conditions with respect to nitrogen and carbon availability, and included the levels of 45 metabolites from 97 *A. thaliana* accessions. Because the models largely encompass the reactions from central carbon metabolism, we expect that the structure of the metabolic network remains unaltered between accessions; under this assumption, the data profiles can be regarded as realizations of the same network. Therefore, we selected the sensors and non-sensors from the two models and repeated the correlation analysis.

In the AraCORE model, we could map 12 sensors and 20 non-sensors, while data were available for 10 sensors and 25 non-sensors in the AraSEED model. The correlation between sensors was significantly higher compared to non-sensors in AraCORE and the AraSEED model (Supplemental Figure 8.1.2). This analysis demonstrated that, under simplifying assumptions about robustness of central carbon metabolism in plants, similar patterns between sensors and non-sensors as in the analysis of single

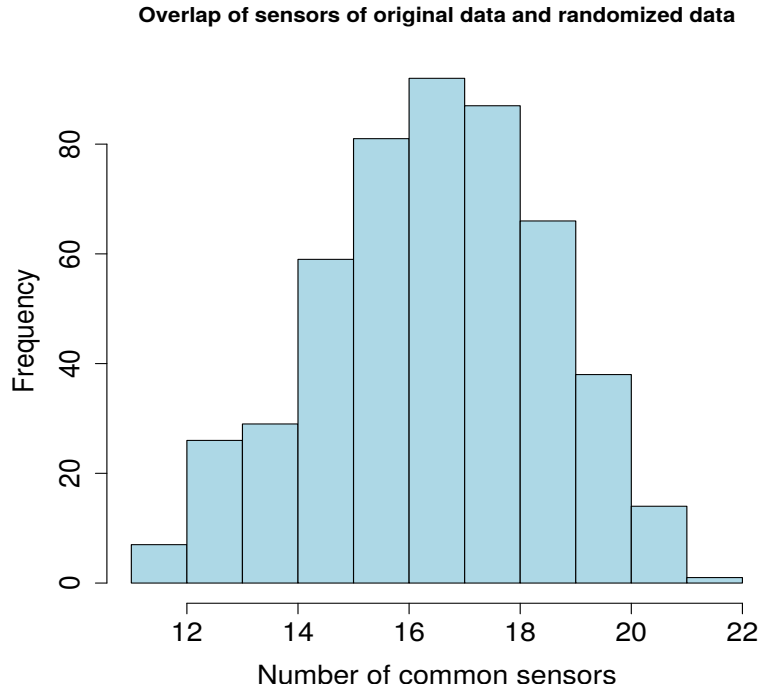

Figure 4.6: **Distribution for the size of the overlap of identified sensors.**
The distribution is obtained after randomizing the reversibility assignment in in the AraCORE model. The x axis displays the number of sensors overlapping with the original analysis. The y axis displays the frequencies of common sensors in 500 shufflings of the reversibility assignment. Originally, 23 sensors were detected.

genotypes can also be found by using data from genetically variable populations.

4.3.3 Analysis of robustness for the observed sensor/non-sensor patterns

To determine if observed pattern of correlations within and between sensors and non-sensors were not artefacts of the used network and could not have resulted by arbitrary grouping of metabolites, we conducted two types of robustness analyses. In the first, we randomized the partition of metabolites into the two classes, while in the second we inspected the effect of the reversibility of reactions considered in the metabolic network.

In the first analysis of robustness, we determined the probability that a random partition of metabolites into same number of sensor and non-sensor metabolites (as in the findings) results in the observed pattern of correlations. To this end, we shuffled the assignment of sensor and non-sensor metabolites 500 times, while keeping their respective total numbers fixed, and determined the data properties, namely, SD and CV as well as correlation for the classes of metabolites and metabolite pairs. This robustness analyses demonstrated that the observed larger correlation of sensors in comparison to non-sensors was statistically significant. In addition, the correlation between sensor and non-sensors was similar to the other two estimated correlations of sensors to sensors and non-sensors to non-sensors. We further supported this finding by the distributions of the three properties in every interval over the considered conditions, which could not be distinguished between the classes of metabolites and metabolite pairs (Supplemental Figure 8.1.3).

It has already been observed that the sensors predicted by the approach we used may change upon

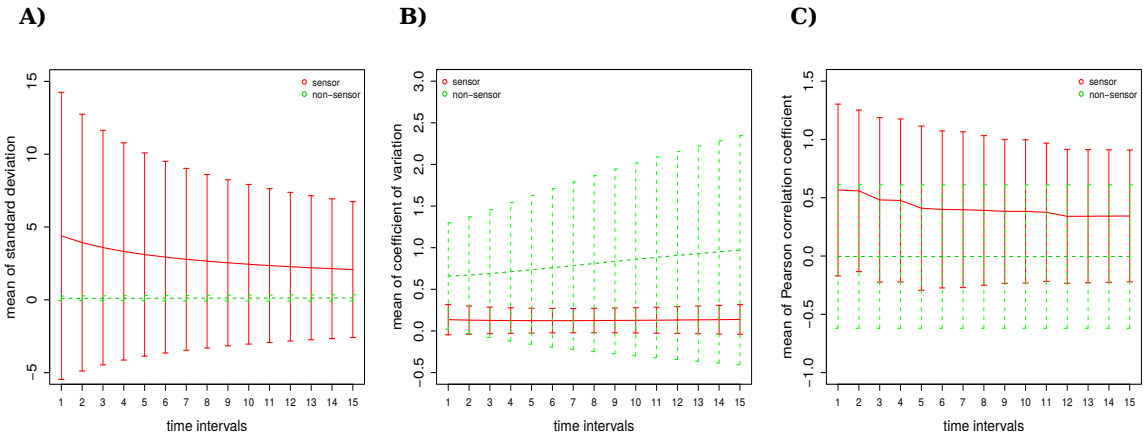

Figure 4.7: **Statistical comparison of the sensors and non-sensors in the kinetic day-time model of plant central carbon metabolism**.
The x axis represents the investigated time interval, from 1 to 15. The y axis represents values for the three statistics, respectively: **A)** SD ; **B)** CV; and **C)** Pearson correlation of sensors and non-sensors. The red line corresponds to the values for the statistics between sensor metabolites, while the green line corresponds to values between non-sensor metabolites. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and non-sensors. Bars represent the range ± 1 SD from the mean value, for five simulations.

alterations of the reaction directionality [Liu et al., 2013]. Therefore, in the second analysis or robustness, we tested the effect of randomizing the reversibility of reactions considered in the model. The 500 randomizations were performed while preserving the number of reversible and irreversible reactions in the network together with the set of metabolites they interconvert. The original sensors in AraCORE consisted of 23 metabolites, whereas after randomization we found between 25 and 42 sensors, of which 11 - 22 (i.e. at least 48%, see Figure 4.6) were also present in the original set of sensor metabolites. Moreover, each of the sensors was identified in at least one randomization. Therefore, the results supported the robustness of the identified sensors and were in line with existing studies, which have pointed out that reversibility of biochemical reactions had a small effect on the identified sensors [Liu et al., 2011]. Similar results were obtained for the second model; here, after randomization, we found between 108 and 153 sensors, of which 57 -90 (i.e. at least 28.79%) were identical with the 198 sensors in the original network. The overlap with the original sensors was lower, compared to the other two models; nevertheless, we could capture in more than half of the permutations, a > 36% overlap. These results were also partly in support of our claim for the robustness of sensors.

4.3.4 Test on kinetic model of central carbon metabolism

To further validate the finding that sensor metabolites are more correlated with each other than non-sensor metabolites, we repeated the analysis with a synthetic data set generated from a medium-size kinetic model of plant central carbon metabolism. The model included the Calvin-Benson-Cycle, triose phosphate transport, sucrose biosynthesis and degradation, starch biosynthesis and degradation, photorespiration, ATP synthesis and the photosynthetic electron transport distributed over five compartments. It comprised 78 metabolites and 112 reactions, representing the largest kinetic model of plant central metabolism to date [Hahn, 1986; Singh and Ghosh, 2006]. This model however does not contain the TCA cycle and the vast majority of amino acids. The reaction rates were modeled according to mass action kinetic (see Supplemental Kinetic Model for the stoichiometric matrix and reaction

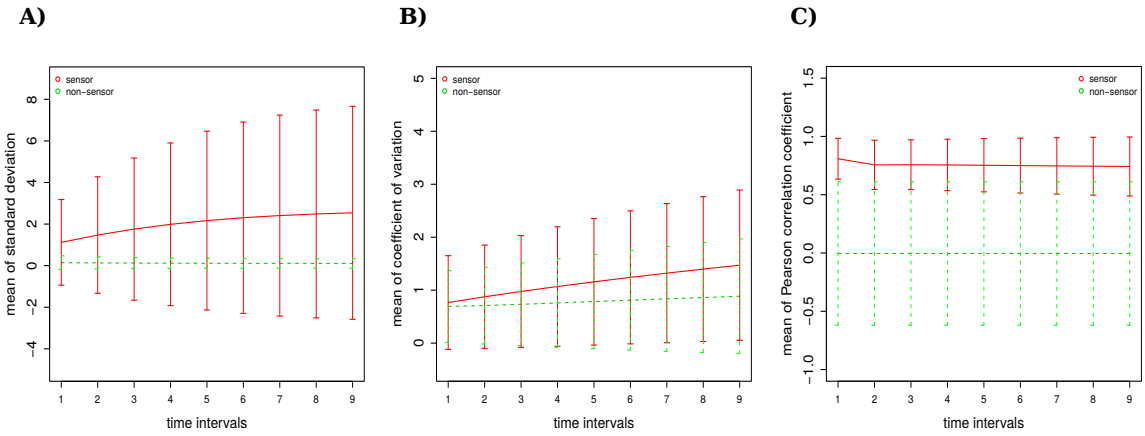

Figure 4.8: **Statistical comparison of the sensors and non-sensors in the kinetic night-time model of plant central carbon metabolism.**
The x axis represents the investigated time interval, from 1 to 9. The y axis represents values for the three statistics, respectively: **A)** SD ; **B)** CV; and **C)** Pearson correlation of sensors and non-sensors. The red line corresponds to the values for the statistics between sensor metabolites, while the green line corresponds to values between non-sensor metabolites. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and non-sensors. Bars represent the range ± 1 SD from the mean value, for five simulations.

parameters in Section 8.1).

In the case, we identified six sensor metabolites, solely using the approach based on the network structure, including 2-oxoglutarate, serine in the mitochondrion, sucrose in the cytosol and the vacuole, as well as hydrogen peroxide (H_2O_2) and ammonia. Based on the simulated data profiles (by varying the initial conditions), we again found that the correlation within sensors was higher than within non-sensor metabolites, for both day and night conditions (Figure 4.7 and Figure 4.8). The SD of the sensors was in both cases higher than for the non-sensors, similar to the results of the AraSEED model. The results of the CV differed between day and night simulations. The day simulation showed a pattern which was comparable to AraCORE (see Figure 4.7), while the night simulations showed similarities to the AraSEED model results (see Figure 4.8). Altogether, the findings from the simulated data profiles from a medium-size kinetic model were in line with the data gathered from experiments, particularly with respect to the observed ordering of correlations within and between groups of metabolites.

4.3.5 Implications of the findings

In this study we demonstrated that metabolites identified as sensors were more correlated than non-sensor metabolites based on data profiles gathered from wet-lab experiments as well as *in silico* simulations. Most of the findings were independently reproduced for two well-curated models of *A. thaliana*. Furthermore, we showed that this was not due to an artifact of the used data by an extensive robustness analysis. By randomly assigning the labels “sensor” and “non-sensor” to the metabolites in the analyzed data set, we demonstrated that the correlation of sensors, non-sensors and between sensors and non-sensors could no longer be observed. Additionally, we tested the influence of reversible reactions in metabolic networks. Importantly, we could reproduce these results on a kinetic model of medium size used for simulating a synthetic data set. Using random but physiologically viable initial conditions for the simulation of day and night cycles, we found the same relationship

between sensor metabolites and non-sensor metabolites as in the Arabidopsis large-scale models.

The identified sensors in all models were metabolites which act as major building blocks of biomass. In the smaller AraCORE, we found cellulose for cell wall synthesis and most of the amino acids for the protein biosynthesis, as well as nucleotides for DNA and RNA replication. These were in agreement with the results of the genome-scale Arabidopsis model, AraSEED, in which in addition to the mentioned metabolite classes, we also identified Coenzyme A and related metabolites playing important role in the TCA cycle [Fatland et al., 2002].

Our results largely depend on the quality of the networks employed. Therefore, we critically investigated the network models used and found that a large number of sensor metabolites were in fact dead-end metabolites, created upon removal of the biomass reactions. By consideration of the respective biomass reaction, the identified metabolites were not dead-end metabolites just in AraCORE.
In AraSEED with a biomass reaction, 4 of the 15 sensors were dead-end metabolites. Investigation of the large-scale metabolic networks used in the study of Liu et al. [2013] showed similar results: In the human RECON1 model, yeast, and *Escherichia coli* models, 57.04%, 76.92%, and 59.81% of the sensors were dead-end metabolites. This is in line with a claim of Liu et al. [2013] that all pure products, i.e. metabolites which do not act as reactants in a single reaction, can serve as sensors. A potential explanation of these high numbers of blocked reactions is that most models contain only an incomplete set of catabolic reactions. Therefore, more metabolites may be predicted as sensors by the approach, as degrading reactions might be missing.

While our empirical tests were built around time-courses within single genotypes, we demonstrated that similar relationships among our predicted sensors could be found in genetic populations of *A. thaliana*. Analogs to these results have also been observed in correlation-based network analysis of metabolic profiles from tomato (*Solanum lycopersicum*) introgression line mapping population, where five amino acids (i.e., glycine, isoleucine, serine, threonine, and valine) were significantly more correlated (average value of 0.84) in comparison to the average correlation between any other measured metabolites [Toubiana et al., 2012, 2015]. Thus, our predicted sensors may be useful to understand the correlations arising in genetically variable populations.

4.4 Conclusion

In this work, we aimed to identify if there are features of the data profiles of sensor metabolites, identified with well-established network-based approaches, that separate them from the rest of the metabolites in a given large-scale plant metabolic network. Methods from observability theory allow computationally feasible identification of sensor metabolites; however, the existing studies have not investigated the extent to which the data profiles of sensors may differ from those of non-sensor metabolites. By employing experimentally and *in silico* generated time-series metabolomics data together with large- and medium-scale structural and kinetic models of Arabidopsis central metabolism [Dall'Osto et al., 2012], we demonstrated that sensor metabolites are, on average, more correlated than non-sensor metabolites across employed models and data sets. Our analyses of robustness further confirmed that these results were due to the position of the sensor metabolites in the network, and complement the implications from other approaches. These correlations tend to persist irrespective of the conditions as long as the underlying functionality of the network, a result of the set of the operational biochemical reactions, remains largely unchanged, as illustrated on data from natural variation. As a

result, our study suggests that relatively few key metabolites could be measured to potentially characterize the entire metabolic network, opening the possibility for applications of targeted metabolite analyses guided by predictions from large-scale models as a means of providing a rapid yet accurate synopsis of the metabolic status of a plant system.

5 Stoichiometric correlation analysis: principles of metabolic functionality from metabolomics data

Publication: Front. Plant Sci. | doi: 10.3389/fpls.2017.02152

Authors: Kevin Schwahn[1,2], Romina Beleggia[3], Nooshin Omranian[1,2,4], Zoran Nikoloski[1,2,4]*

Affiliations: [1]Systems Biology and Mathematical Modeling Group, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, Potsdam-Golm, Germany

[2]Bioinformatics Group, Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24-25, Potsdam-Golm, Germany

[3]Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria, Centro di Ricerca per la Cerealicoltura e le Colture Industriali (CREA-CI), 71122 Foggia Italy

[4]Center of Plant Systems Biology and Biotechnology (CPSBB), Plovdiv, Bulgaria

Contact: *nikoloski@mpimp-golm.mpg.de

Abstract

Recent advances in metabolomics technologies have resulted in high-quality (time-resolved) metabolic profiles with an increasing coverage of metabolic pathways. These data profiles represent readouts from often non-linear dynamics of metabolic networks. Yet, metabolic profiles have largely been explored with regression-based approaches that only capture linear relationships, rendering it difficult to determine the extent to which the data reflect the underlying reaction rates and their couplings. Here we propose an approach termed Stoichiometric Correlation Analysis (SCA) based on correlation between positive linear combinations of log-transformed metabolic profiles. The log-transformation is due to the evidence that metabolic networks can be modeled by mass action law and kinetics derived from it. Unlike the existing approaches which establish a relation between pairs of metabolites, SCA facilitates the discovery of higher-order dependence between more than two metabolites. By using a paradigmatic model of the tricarboxylic acid cycle we show that the higher-order dependence reflects the coupling of concentration of reactant complexes, capturing the subtle difference between the employed enzyme kinetics. Using time-resolved metabolic profiles from *Arabidopsis thaliana* and *Escherichia coli*, we show that SCA can be used to quantify the difference in coupling of reactant complexes, and hence, reaction rates, underlying the stringent response in these model organisms. By using SCA with data from natural variation of wild and domesticated wheat and tomato accessions, we demonstrate that the domestication is accompanied by loss of such couplings in these species. Therefore, application of SCA to metabolomics data from natural variation in wild and domesticated populations provides a mechanistic way to understanding domestication and its relation to metabolic networks.

5.1 Introduction

Metabolomics profiling technologies are increasingly used for phenotyping of biological systems to understand the contribution of metabolism to complex phenotypes, including growth and diseases [Schauer and Fernie, 2006; Sumner et al., 2003; Kaddurah-Daouk et al., 2008]. They have been used to assess the relative and absolute levels of different metabolites after perturbation or over time [Fiehn et al., 2000]. The resulting metabolic data profiles manifest the joint effect of the rates of multiple biochemical reactions interrelated in metabolic networks. Reaction rates are themselves subjected to different types of regulation, often carried out by altering the concentration of metabolites [Koshland, 1970].

Regulation of reaction rates is necessary to ensure that the activities attributed to different parts of the system are coordinated. The simplest way to capture the coordination of reactions rates is through their coupling, whereby the ratio of the reaction rates is maintained in a narrow range [Millard et al., 2017], resulting in high positive correlation values between the coupled reaction rates over different experiments (e.g., environments). The principle questions in analyzing the data from metabolomics technologies are then to determine the extent to which the metabolite levels reflect the coupling of the underlying biochemical reactions as well as any differences in these characteristics between experimental scenarios (e.g. comparison of genotypes or treatments).
Despite the apparent non-linearites due to the metabolic structure and regulation, metabolic data profiles are usually analyzed by regression-based approaches that can only capture linear relationships.

Ever since the seminal work of Vance et al. [2002], which used partial correlations to analyze the dependence between metabolites and reconstruct the reactions in which they participate, the existing analyses of metabolic data profiles rely on applying various similarity measures to given metabolic profile [Çakır et al., 2009; Krumsiek et al., 2016]. Since correlation, like other similarity measures, results in bilateral relationships between metabolites, the resulting metabolite-metabolite relationships have been represented and analyzed in the framework of metabolic correlation network analysis (MCNA) [Toubiana et al., 2013]. This has led to the usage of MCNA to compare data from different scenarios based on the concept of differential networks [Chen et al., 2009; Ideker and Krogan, 2012]. However, the principle question about coupling of biochemical reactions reflected in the metabolic profiles remains unresolved.

Assuming random fluctuations around a given steady state, metabolic correlations have been related to the Jacobian of the system of ODEs that describe the change in metabolite concentrations [van Kampen, 2007]. In a series of studies, this relation has been employed for reconstructing the Jacobian of simplified metabolic networks and for comparison of different treatments [Steuer et al., 2003; Sun et al., 2015; Nägele et al., 2016]. While this approach places metabolic correlations on strong theoretical basis, it is not applicable for analysis of instationary data. In another network-driven approach [Hackett et al., 2016], metabolic profiles have been fitted to steady-state compatible fluxes (extracted under optimality assumption of the flux balance analysis [Orth et al., 2010]) with different functional form for the reaction rates $v(x,k)$. This approach has allowed the elucidation of novel regulators of reaction rates.

Here we take a principally different approach motivated by biochemically reasonable assumptions which often hold in realistic biological scenarios. Since biological systems sense and respond to environmental perturbations, they achieve normal functionality in face of these perturbations. To this end, various feedbacks and mechanisms based on network structure have evolved to maintain coupling of reaction rates. Based on this idea and under the assumption that elementary biochemical reactions can be modeled via mass action kinetics (without neglecting the effect of enzymes), here we propose a novel means to analyze metabolic profiles based on the concept of constrained maximal correlation coefficient. We use this approach to analyze and characterize the role of metabolites in a network that captures the reaction rate coupling. First, by using a paradigmatic model of the tricarboxylic acid (TCA) cycle, we investigate the effect from departures of the assumption of mass action on the identified reaction coupling and couplings of reactant complexes. We then show that Stoichiometric Correlation Analysis (SCA) can be employed to perform cross-species comparison of the TCA cycle and amino acid synthesis pathways. In addition, we demonstrate that the proposed approach can be used to mechanistically understand the agronomically important process of domestication, here, in the case of wheat as well as in tomato and strawberry.

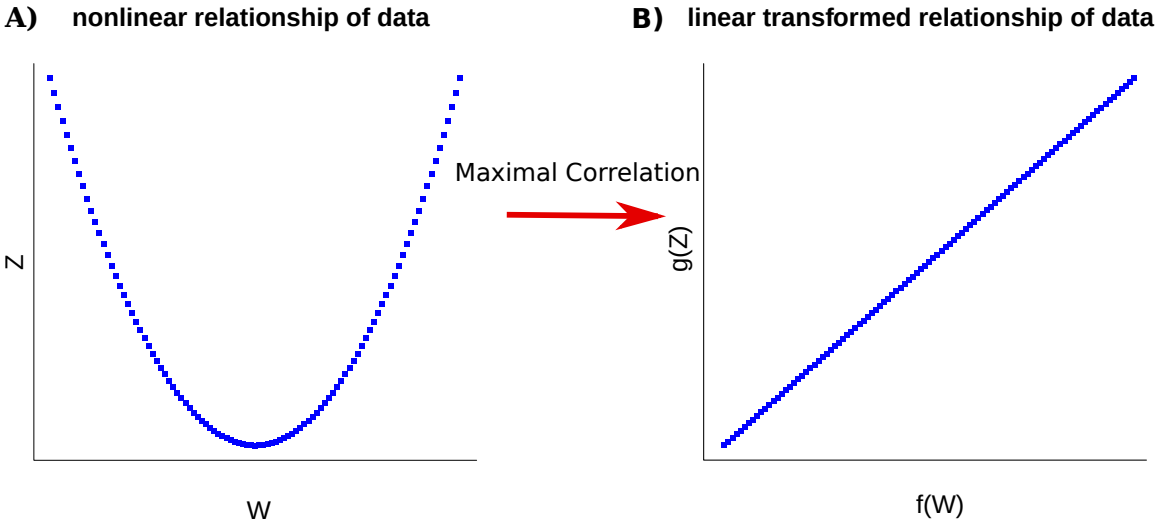

Figure 5.1: **Representation of the Maximal Correlation**
A) The relationship of the variables W and Z is nonlinear. **B)** Employing maximal correlation finds the functions f and g. These allow the transformation of the data and capture the underlying relationship between the variables W and Z.

5.2 Materials and Methods

5.2.1 Description of the approach with the underlying assumptions and principles

Maximal correlation

Modern applications, particularly in computational biology, often consider a large number of variables involved in nonlinear (pairwise) relationships. The maximal correlation coefficient, ρ, between a pair of random variables W and Z, introduced by Gebelein [1941] and already extensively studied by Lancaster [1957] and Rényi [1959], is defined as:

$$\rho = sup\left\{\frac{cov(f(W),g(Z))}{\sqrt{(f(W))V(g(Z))}}|V(f(W))>0, V(g(Z))>0\right\}, \tag{5.1}$$

where the supremum is taken over all functions f of W and g of Z with finite variances, i.e., $V(f(W))>0$ and $V(g(Z))>0$. Maximal correlation then infers (non-linear) transformations of two random variables by maximizing their pairwise correlation (see Figure 5.1 for illustration). We note that W and Z are independent if and only if $\rho = 0$, relating maximal correlation to mutual information (see Introduction in Section 5.1).

There exist efficient algorithms to compute maximal correlation for both discrete [Breiman and Friedman, 1985] and continuous [Lancaster, 1957] random variables. Direct application of these algorithms for calculation of maximal correlation to time-resolved metabolic profiles is hampered since: (1) metabolic profiles are quantitative (i.e. continuous variables), as they capture the content of metabolic pools in biological systems; therefore, any decision to move to a range of values (e.g. small, medium,

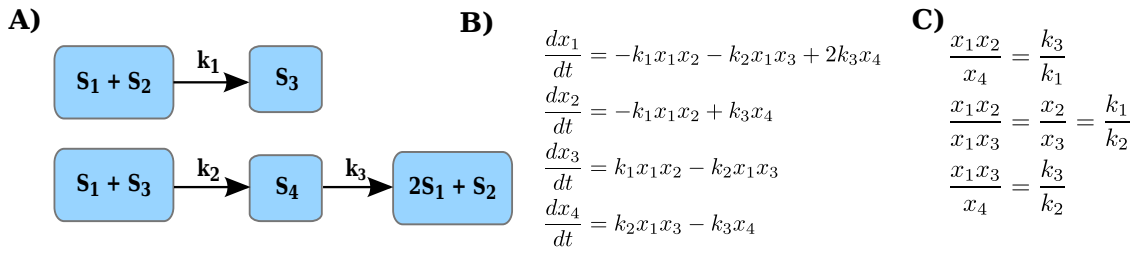

Figure 5.2: **Illustration of reaction couplings**.
A) Network with four components, $S_1 - S_4$, and three reactions with rate constants $k_1 - k_3$; **B)** A system of ODEs with mass action kinetics describing the change in concentration of each of the four components. **C)** Couplings of reaction rates by invoking the steady-state assumption for the system of ODEs in B.

large, as it is done in discretization), will lead to drastic simplification, and (2) time-resolved metabolic profiles include relatively few time points, rendering the calculation of maximal correlation based on contingency table challenging (e.g. Nguyen et al. [2014] analyzed maximal correlation with at least 100 data points which is still not available for metabolomics data).

Stoichiometric Correlation Analysis and the principle of metabolic network robustness

Here we define a constrained version of the maximal correlation coefficient which is motivated by modeling of metabolic networks and the principles of their operation. A metabolic network is a collection of metabolites and biochemical reactions through which they are transformed and/or exchanged with the environment. For instance, the network on Figure 5.2A transforms four metabolites, S_1 to S_4 via three reactions. Each reaction takes a non-negative linear combination of reactants metabolites, called substrate complex, and transforms it into a product complex, i.e. a non-negative combination of product metabolites. The coefficients in the non-negative linear combination denote the stoichiometry with which a metabolite enters a reaction as a substrate and/or product. For instance, in Figure 5.2A, $S_1 + S_2$ is the substrate complex of reaction r_1 and $2S_1 + S_2$ is the product complex of reaction r_3. The difference between the stoichiometry of the product and substrate complexes defines a reaction vector stoichiometry gathered in the stoichiometric matrix N. In other words, the entry α_{ij} of the stoichiometric matrix N contains the molarity (integer number) with which metabolite i is involved as a substrate or product in the reaction j [Heinrich and Schuster, 1996].

The change in the levels of n metabolite $x_1, \ldots, x_n$ can then be described by an ordinary differential equation (ODE), $\frac{dx}{dt} = N * v(x,k,t)$ where N denotes the stoichiometric matrix with dimensions $m \times n$, with m the number of metabolites and n the number of reactions, v denotes the reaction rate functions, x, the concentrations of the considered metabolites, k, the parameters on which the reaction rates depend, and t stands for time [Nägele et al., 2016]. Even in the simple case of mass action kinetics for a network of bimolecular reactions, the reaction rates, gathered in the time-dependent vectors, $v(x,k,t)$, are described by a non-linear function [Horn and Jackson, 1972]. Metabolic reactions usually are not spontaneous and are catalyzed by enzymes. Every enzymatic reaction can in turn be divided into elementary reactions. Elementary reactions consider the formation and dissociation of enzyme-substrate complexes and provide the possibility for modeling variety of regulatory mechanisms [Segal, 1975]. Elementary reactions can be effectively modeled with mass action, since they can be cast to explicitly consider the action of the enzyme (as in the derivation of the Michaelis-Menten kinetic). This was the approach taken in the large-scale model of *Escherichia coli* [Khodayari and

Maranas, 2016] and some of the subsystems in the models of photosynthesis [Arnold and Nikoloski, 2011]. Therefore, due to the combined effect of multiple reactions and their regulation, metabolic data profiles can be regarded as observations from non-linear dynamics of metabolic networks.

Let x_i denote the concentration of a substrate component S_i (i.e. metabolite or enzyme). The rate of reaction j with a substrate complex $\sum_i \alpha_{ij} S_i$ with $\alpha_{ij} > 0$ where α_{ij} is the stoichiometry with which S_i enters the substrate complex of reaction j, under mass action kinetics is then expressed as $k_j \prod_i x_i^{\alpha_{ij}}$, where k_j denotes a rate constant. For instance, the rate of reaction with rate constant k_1 in Figure 5.2A is given by $k_1 x_1 x_2$, since S_1 and S_2 enter this reaction as substrates, each with stoichiometric coefficients of one; similarly, the rate of reaction with rate constant k_3 is given by $k_3 x_4$, since S_4 enters the reaction as a substrate with a stoichiometric coefficient of one.

To arrive at our approach termed Stoichiometric Correlation Analysis (SCA) we rely on the observation that metabolic networks, as part of inter-related cellular systems (e.g. transcription, translation, and signaling), operate towards providing robust functionality [Wilson, 2013; Kitano, 2007]. We translate the robust functionality in the ability to ensure coupling of reaction rates [Millard et al., 2017]. To formalize SCA, we provide the following definitions:
Definition 1: Two elementary reactions, p and q, have coupled rates under mass action kinetics if for any steady-state concentration of the participating components, gathered in x,

$$\frac{k_p \prod_i x_i^{\alpha_{ip}}}{k_q \prod_i x_i^{\alpha_{iq}}} = \frac{k_p}{k_q} \prod_i x_i^{\alpha_{ip}-\alpha_{iq}} = \gamma_{pq}, \text{ where } \gamma_{pq} \text{ is a constant.}$$

For instance, at any steady state for the network in Figure 5.2A, whereby the equations in Figure 5.2B all equal 0, i.e. $\frac{dx_i}{dt} = 0$, reactions r_1 and r_2, r_1 and r_3 , as well as r_3 and r_2 have coupled rates (see Figure 5.2C). We note that the same definition can be extended to hold in states which are not necessarily equilibrium points, allowing the treatment of time-series data. We would like to note that the coupling of reaction rates may lead to coupling of component concentrations which are not apparent by directly inspecting the reaction networks. For instance, due to the coupling of reactions r_1 and r_2, the concentration of components x_1 and x_2 are also coupled, i.e., are proportional to each other.

Since the non-zero stoichiometric coefficients α_{ij} are integers in the set $I = \{1, \ldots, 4\}$ [Basler et al., 2012], given two disjoint sets U_p and U_q of random variables denoting the data profiles for the metabolites, we next define the stoichiometric correlation.
Definition 2: Given two disjoint sets of random variables U_p and U_q, denoting two sets of metabolic profiles, the stoichiometric correlation is given by:

$$sup \left\{ \frac{cov(f(U_p), g(U_q))}{\sqrt{V(f(U_p))V(g(U_q))}} | V(f(U_p)) > 0, V(g(U_q)) > 0 \right\}, \tag{5.2}$$

with

$$f(U_p) = \sum_{i=1}^{|U_p|} \beta_{ip} log(x_i), \beta_{ip} \in I$$

and

$$g(U_q) = \sum_{i=1}^{|U_q|} \eta_{iq} log(x_i), \eta_{ip} \in I.$$

If U_p and U_q include the random variables corresponding to the metabolite levels in the substrate complexes of reaction p and q, respectively, the proposed definition of the stoichiometric correlation is a direct consequence of Definition 1, where the functions $f(U_p)$ and $g(U_q)$ are the logarithm of the rate of the reactions p and q under mass action kinetics, respectively. The presence of coupled rates in mass action for reactions p and q, after taking the logarithm, leads to the stoichiometric correlation of value one for U_p and U_q. This observation pinpoints the main principle on which SCA relies.

If there exist multiple vectors β and η, yielding the same value of the stoichiometric correlation, we consider the one of smallest magnitude $\left\|\beta+\eta\right\|_2$. Therefore, stoichiometric correlation can be regarded as constrained maximal correlation, where the constraints pertain to the limited set of values that the entries of β and η are allowed to take following the stoichiometry of reactants. The transformation used in the constrained maximal correlation is explicitly non-linear, since the functions f and g involves logarithms.

Clearly, the reverse direction also holds and can be used to draw hypotheses about the couplings in reaction rates and substrate complexes in a given metabolic network. To this end, we focus on the statistically significant stoichiometric correlations larger than a threshold value of 0.8 (to account for effects of noise and small deviations from coupling of reaction rates, per Definition 1). Note that since U_p and U_q are disjoint sets of random variables denoting the data profiles of metabolites, the entries β of η and are positive. For instance, given several steady-state measurements for the components in the network on Figure 5.2A, the stoichiometric correlations with $U_p = \{S_1, S_2\}$ and $U_q = \{S_4\}$ is one with coefficients in β and η equal to one. Similar conclusions can be drawn for all components involved in the coupled reaction rates given in Figure 5.2C. The two definitions provide the basis for SCA: Since majority of reactions in real-world metabolic networks are mono- or bi-molecular (i.e. include one or two substrates), we determine the stoichiometric correlation, per Definition 2, between any two disjoint subsets of random variables of cardinality at most two. The implementation can either be achieved by: (1) solving a non-linear program with constraints for the coefficients $\beta, \eta \in I$ or (2) generating all subsets of at most two variables with different contribution due to stoichiometry, and determining the Pearson correlation coefficient only between the disjoint subsets. Since the number of available metabolic profiles from time-resolved studies usually does not exceed 100, the second alternative can be efficiently implemented with appropriate parallelization (see the code in Schwahn et al. [2017b]). The significance of the stoichiometric correlation can be readily estimated by permutation tests after adjusting for multiple hypotheses testing.

5.2.2 Implementation of SCA

Given a data set of n metabolites over c samples (i.e. each representing a particular time point in an environment), we implemented SCA by determining: (1) the Pearson correlation $r(log(x_i), log(x_j))$, for all couples $1 \leq i \neq j \leq n$ of metabolic profiles, (2) the values for $a, b \in \{1,2,3,4\}$ that maximize the Pearson correlation between $a\,log(x_i)+b\,log(x_j)$ and $log(x_k)$ for every triple of metabolic profiles, (3) the values for $a, b, c, d \in \{1,2,3,4\}$ that maximize the Pearson correlation between $a\,log(x_i)+b\,log(x_j)$ and $c\ log(x_k)+d\ log(x_l)$ for every quadruple of metabolic profiles. In addition, we determined the statistical significance for each of the maximum correlations. We used the R package Hmisc [Harrel, 2015] to calculate the correlation and associated p-values. In addition, we adjusted the p-values using Benjamini-Hochberg multiple hypotheses testing correction. We considered stoichiometric correlations with adjusted p-values below $\alpha = 0.05$ as significant.

The code and one example can be found on GitHub: https://github.com/KSchwahn/Stoichiometric-correlation [Schwahn et al., 2017b].

5.2.3 Models

Metabolite levels were simulated with three different models using Michaelis-Menten kinetics [Singh and Ghosh, 2006], mass action kinetics and extended mass action kinetics with metabolite-enzyme complexes [Khodayari et al., 2014]. The Michaelis-Menten based model contains 11 reactions and 12 metabolites and simulates the metabolite levels within the TCA cycle of *E. coli* growing on glucose. The synthetic reaction (SYN) and the biomass metabolite (biosyn) were removed, as a comparable reaction and metabolite were not present in the other two analyzed models. The modified Michaelis-Menten model contains therefore 11 metabolites and 10 reactions. All kinetic parameters remained unchanged. The mass action based models contain the TCA cycle of the *E. coli* model of Khodayari et al. [2014]. The solely mass action based model contains 23 metabolites and 22 reactions after splitting each reaction into a forward and backward reaction. The second model includes the simulation of metabolite-enzyme complexes based on mass action kinetics. The model contains 114 irreversible reactions and a total of 80 metabolites, enzymes and metabolite-enzyme complexes.

The change of concentration was simulated with each model over a time course of 1,280 minutes. The initial concentration of the metabolites, metabolite-enzyme complexes and enzymes was randomly assigned for each of the 10 repetitions from the range of the minimum and maximum metabolite concentration reported in Khodayari et al. [2014]. The same set of 11 metabolites, present in each model, was then used for the calculation of stoichiometric correlations with the SCA approach. All simulations were performed in MATLAB 2015a [The MathWorks, 2015].

5.2.4 Metabolic data profiles

We applied SCA to several publicly available metabolomics data sets, including metabolic profiles from *Arabidopsis thaliana* obtained from Caldana et al. [2011] and *Escherichia coli* from Jozefczuk et al. [2010]. The first consists of data profiles of 92 metabolites over eight conditions measured over 22 time points with 6 replicates each (light and dark at 4°, 21°, and 32°C, as well as low light at 21°C and high light at 21°C; high light was discarded as it contains less time points), while the second includes 196 metabolites over five conditions measured over 12 time points with three biological replicates per time point and three technical replicates each (cold stress, heat stress, oxidative stress, lactose and control condition).

We also used the metabolomics data from a recent evolutionary metabolomics study [Beleggia et al., 2016]. The study identified and quantified 51 metabolites from nine compound classes in the three taxa of wheat, namely, wild emmer, emmer, and durum wheat. The metabolites were measured in kernels of 12 accessions from wild emmer, 10 from emmer, and 15 accessions from durum wheat, whereby the measurements contain three biological replicates with three technical replicates each. Like the other data sets used here, the metabolic profiles in the wheat taxa were assessed by gas chromatography-mass spectrometry. To allow comparability between taxa, we used only the 22 metabolites, from four compound classes, which were detected across all accessions.

Moreover, we included metabolomics data from six different tomato species, namely *Solanum chmielewskii*, *Solanum habrochaites*, *Solanum lycopersicum*, *Solanum pimpinellifolium*,

Solanum neorickii, and *Solanum pennellii* [Schauer et al., 2005]. We consider data from the ripe fruit in this study. These data are included to further test our assumption about the effect of domestication on reaction coupling. Altogether, we compare 43 metabolites for the tomato data. The *S. lycopersicum* metabolomics measurements were obtained from the study of Schauer et al. [2006] and contain 108 replicates from the year 2001 and 84 replicates from the year 2003, whereas the remaining data were obtained from Schauer et al. [2005] and contain six replicates for each of the five species.

To have a more comprehensive comparative analysis pertaining to domestication, we included further data of wild strawberry accessions (*Fragaria vesca*) and a domesticated strawberry species (*Fragaria ananassa*) [Ulrich and Olbricht, 2013]. Overall, 19 different metabolites had complete measurements to be included in the analysis. The data set contains measurements from 32 samples of *F. vesca* and 10 samples of *F. ananassa*. This data set in comparison to the other data set contains specifically the volatile organic compounds extracted from the strawberry fruits.

5.3 Results and Discussion

5.3.1 Stoichiometric Correlation Analysis with a paradigmatic model of the TCA cycle

From the derivation of our SCA, it follows that the findings based on the constrained correlation of metabolic data profiles reflect the apparent couplings of elementary reaction rates, assumed to obey mass action kinetic. In addition, the findings reflect the additional couplings which cannot be directly related to reaction rates but are direct consequence of them (e.g., components S_1 and S_2 in the network on Figure 5.2A are coupled due to the coupling of the rates of reactions r_1 and r_2). We note that every enzymatic reaction $\sum_i \alpha_{ij} S_i \rightarrow \sum_i \alpha'_{ij} S_i$ (α_{ij}/α'_{ij} are the stoichiometry with which S_i enters the substrate/product complex of reaction j, respectively) can be rewritten to include the action of an enzyme $\sum_i \alpha_{ij} S_i + E \leftrightarrows SE \rightarrow \sum_i \alpha'_{ij} S_i + E$ (E denotes the enzyme and SE the substrate-enzyme complex), so that the elementary reaction can be still modeled with mass action kinetic. Therefore, SCA can also include the effect of enzyme action. However, while this approach provides a way to model Michaelis-Menten kinetic which accounts for enzyme saturation, it does not explicitly consider the Michaelis-Menten form for the reaction kinetic.

To investigate the effects of the departure from the mass action kinetic for the considered reactions (with and without accounting for enzyme action), we considered three models of the tricarboxylic acid (TCA) cycle. All three models include the same metabolites, and differ only with respect to whether or not they include the effect of enzyme action and if they use mass action kinetic or the more involved functional forms of the Michaelis-Menthen kinetic. All reactions are considered reversible, and they are split into irreversible reactions in the cases in which mass action kinetic was employed. We used the TCA cycle model embedded in the kinetic model of *E. coli* [Khodayari et al., 2014]. There are two parameterized variants for this model, one that includes mass action kinetic without enzyme action, and another one which explicitly considers the formation of substrate-enzyme complexes. In addition, we used a model of the TCA cycle with reversible Michaelis-Menten kinetic of the reaction rates [Singh and Ghosh, 2006].

To conduct the comparative analysis, we simulated the models metabolite concentrations with physiologically relevant randomly chosen initial values. The simulation time ranged from 0 to 1,280 minutes

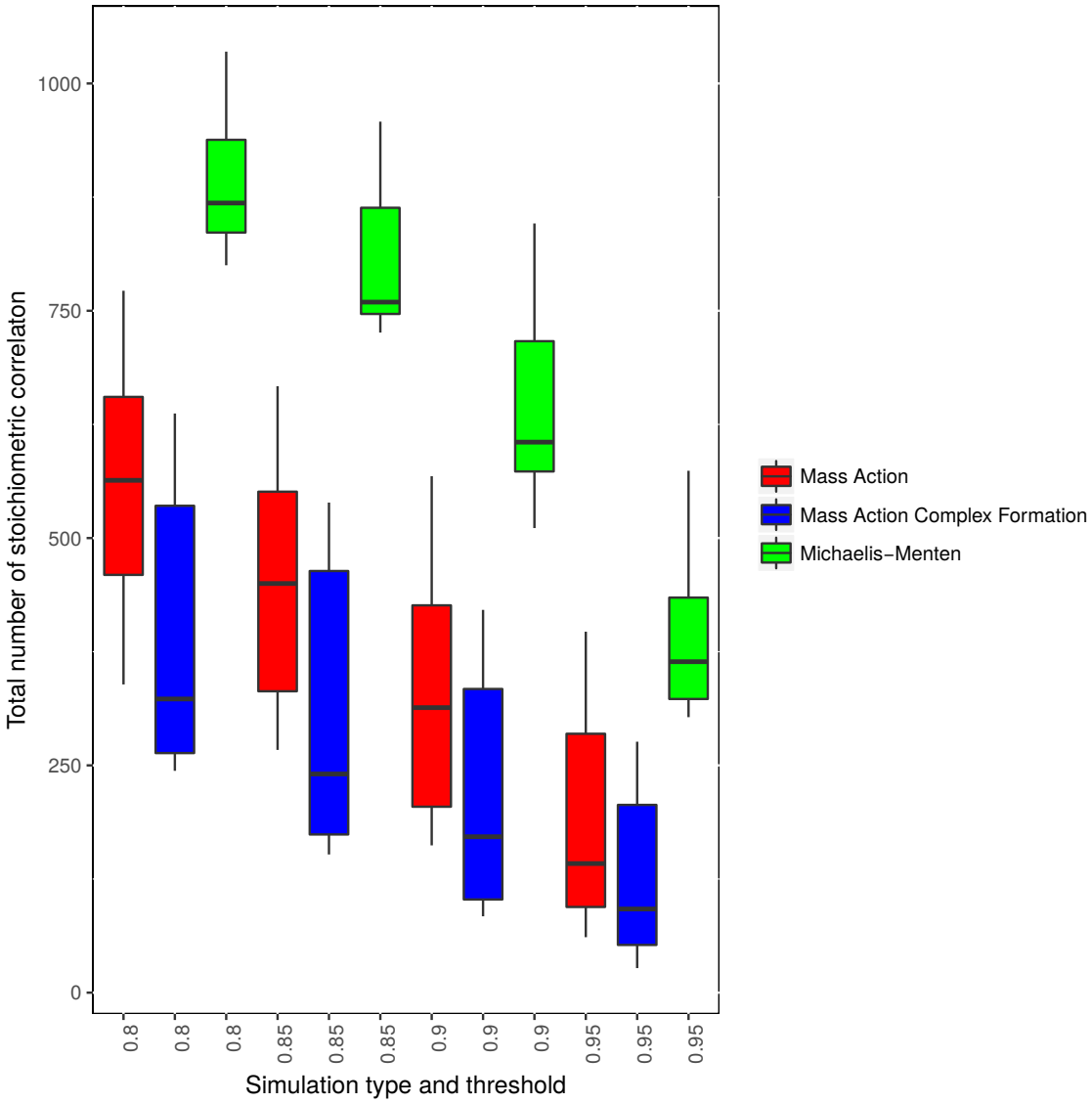

Figure 5.3: **Distribution of the number of stoichiometric correlations for three models of the TCA cycle**
Shown are the distributions of the total number of stoichiometric correlations at four thresholds 0.8, 0.85, 0.9 and 0.95. The distributions for the mass action simulation are shown in red, the distributions for the substrate-enzyme complex mass action simulation are shown in blue, whereas the Michaelis-Menten simulation of the TCA cycle is shown in green.

Table 5.1: Overview of the number of significant stoichiometric correlations at the considered thresholds for metabolic profiles of the stringent response in *E. coli* and *A. thaliana*. The total number of stoichiometric correlations is divided into three groups based on whether they involve pairs, triples, or quadruples of metabolites. Additionally, the number of significant Pearson correlations found in the data set is shown.

		Stoichiometric Correlation				Pearson Correlation
Threshold	Organism	Total	Pairs	Triples	Quadruples	Pairs
0.80	*A. thaliana*	3419	13	579	2827	24
	E. coli	3301	9	517	2775	10
0.85	*A. thaliana*	2500	8	398	2094	18
	E. coli	1921	6	285	1630	7
0.90	*A. thaliana*	1821	6	285	1530	15
	E. coli	597	1	76	520	2
0.95	*A. thaliana*	1137	6	188	943	7
	E. coli	2	0	0	2	0

and metabolite concentrations were obtained at 21 time points identical to those used in the study of Caldana et al. [2011] (which we employ later in the empirical analysis). The simulated metabolite concentrations were used to calculate the stoichiometric correlations for 11 metabolites for each simulation and model separately. The distribution of the total number of stoichiometric correlations over 10 repetitions of the procedure is shown in Figure 5.3, and all stoichiometric correlations (pairs, triples and quadruples) are provided in Supplemental Table 8.2.1.

We found that the total number of stoichiometric correlations between the models with mass action kinetic was more similar with the increase in the considered threshold. In fact, at a threshold of 0.95, the distributions of the total number of stoichiometric correlations between the mass action models with and without the consideration of enzyme action largely overlapped. However, the consideration of reversible Michaelis-Menten kinetic results, on average, in at least three-fold increase in the total number of stoichiometric correlations (see Figure 5.3). These findings were supported by the results of the empirical cumulative distribution function (see Figure 5.4). The distribution of the Michaelis-Menten simulations are shifted to the right and show a higher proportion of correlations above 0.8. In addition, we report the quintiles of the correlation values in Supplemental Table 8.2.2.

Therefore, in the case of the TCA cycle models, we concluded that the findings from the assumption that the network is composed of elementary reactions modeled with mass action do not differ upon consideration of enzyme action. In these cases, the couplings corresponding to the stoichiometric correlations reflect the underlying reaction couplings. In contrast, the usage of Michaelis-Menten kinetic results in a considerably larger number of stoichiometric correlations, which cannot be brought in direct correspondence to the coupling of reaction rates and are challenging to mechanistically explain.

5.3.2 SCA demonstrates differences in the stringent response between *E. coli* and *A. thaliana*

The stringent response is one of the most important regulatory systems used by bacteria to adapt to environmental stresses. Upon sensing the environmental change, like nutrient limitation, the organism starts a series of reactions to redirect its metabolic fluxes. The stringent response is mediated by guanosine 3',5'-bis(pyrophosphate) (ppGpp) whose level is controlled by two enzymes, RelA and SpoT

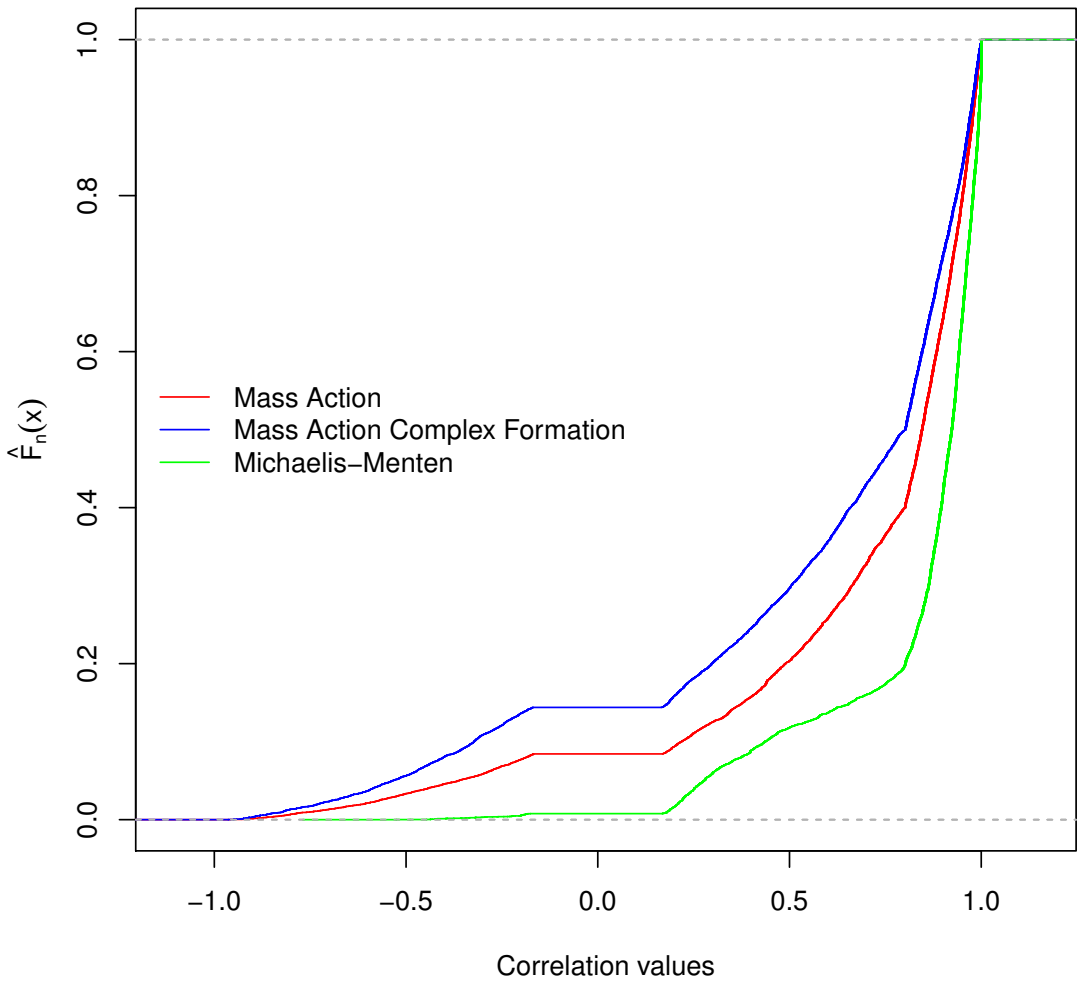

Figure 5.4: **Empirical cumulative distribution function of stoichiometric correlations for three models of the TCA cycle.**
Shown is the Empirical cumulative distribution of the total number of stoichiometric correlations of simulations of the TCA cycle. The distribution for the mass action simulation is shown in red, the distribution for the substrate-enzyme complex mass action simulation is shown in blue, whereas the Michaelis-Menten simulation is shown in green.

[Traxler et al., 2008]. ppGpp has overall a large influence on several metabolic pathways and transcription and translation [Mizusawa et al., 2008; Gallant, 1979]. The effect in metabolism involves the pathways of nucleotides, glycolytic intermediates, carbohydrotes, lipids and fatty acid synthesis. It has been reported that there is evidence that the stringent response is evolutionary conserved from bacterial to photosynthetic bacterial to higher plants [Sugliani et al., 2016]. Four homologues of these RelA and SpotT have been identified in *A. thaliana*, and their role in green tissues and flower development has been well characterized [Masuda et al., 2008; Mizusawa et al., 2008]. Since all these plant proteins are targeted to the chloroplast, it has been suggested that they control the stringent response in photosynthesizing organisms through mechanisms that may mimick those in bacteria. However, it remains unclear to what extent the molecular role of the homologues in *A. thaliana* are equivalent to those in *E. coli*.

To help answer this question, we analyzed the set of metabolites from the TCA cycle and the amino acid synthesis pathways from the two model organisms. We used these metabolites since ppGpp controls transcription and translation, which is ultimately reflected in the levels of amino acids. Moreover, a comparative analysis between the two organisms is only meaningful for the same set of metabolites. Altogether, we used the publicly available data profiles of three metabolites from the TCA cycle (i.e. malate, succinate, and fumarate) as well as 16 amino acid measured over seven and five conditions in *A. thaliana* and *E. coli* (see Section 5.2: Materials and Methods).

The degree of coupling for metabolite S can be defined as the number of stoichiometric correlations above a given threshold τ in which the metabolite S participated. Based on the derivation of SCA, a higher degree of coupling on the same set of metabolites then implies maintenance of more coupled reaction rates over a set of studied conditions in one organism in comparison to another. We considered the significant stoichiometric correlations (p-value ≤ 0.05, Benjamini-Hochberg corrected) 0.8, 0.85, 0.9, and 0.95, and compared them to classical Pearson correlations (Table 5.1). The quintiles of correlation values were additionally reported in Supplemental Table 8.2.2.

For the purpose of comparison, at all threshold values and in both species, we observed a decrease on the number of significant stoichiometric correlations for pairs of metabolites, compared to Pearson correlation (i.e. $\left|U_p\right| = \left|U_q\right| = 1$). The reduction in the number of significant correlations for metabolite pairs can be explained by the monotonic transformation of metabolite profiles. We would like to emphasize that the result does not suggest metabolites are linearly related, which would be contrast to what is expected from mechanistic understanding of metabolism.
However, SCA allows the analysis of stoichiometric correlations due to triples and quadruples of metabolites, which provides information about the presence of non-linear relationships via the couplings of reaction rates. For all considered thresholds, applying SCA with the *E. coli* data set resulted in a smaller number of stoichiometric correlations on triples and quadruples than the data set of *A. thaliana* (Table 5.1). For instance, at a threshold of $\tau = 0.85$, *E. coli* yielded 285 significant stoichiometric correlations due to triples while *A. thaliana* resulted in 398 such correlations; similarly, *A. thaliana* contained three-fold the number of stoichiometric correlations resulting from quadruple at $\tau = 0.9$ in comparison to *E. coli*. Therefore, based on these results we concluded that there was a stronger coupling of reaction rates of *A. thaliana* in comparison to *E. coli* during the stringent response.

Additionally, we can investigate overlapping pairs, triples and quadruples for each threshold. The small similarity of SCA findings was reflected in 65 and 442 stoichiometric correlations due to triples and quadruples, respectively, shared between the two species at a threshold value of 0.8 (see Supplemental Table 8.2.3). In line with this observation, the participation of metabolites in the stoichiometric

Table 5.2: Overview of number of significant stoichiometric correlations at different thresholds for the considered tomato and strawberry species. The total number of stoichiometric correlations is divided into three groups based on whether they involve pairs, triples, or quadruples of metabolites. Additionally, the number of significant Pearson correlations found in the dataset is shown.

		Stoichiometric Correlation				Pearson Correlation
Threshold	Organism	Total	Pairs	Triples	Quadruples	Pairs
0.80	Tomato wild type	19519	8	1245	18266	27
	M82	23571	15	1824	21732	12
	F. vesca	1346	6	204	1136	6
	F. ananassa	2374	12	433	1929	5
0.85	Tomato wild type	9291	5	588	8698	20
	M82	9539	5	688	8846	4
	F. vesca	504	2	73	429	3
	F. ananassa	2075	10	366	1699	5
0.90	Tomato wild type	3741	3	255	3483	11
	M82	1493	1	112	1380	0
	F. vesca	135	1	22	112	1
	F. ananassa	1153	2	185	966	1
0.95	Tomato wild type	818	1	76	741	3
	M82	21	0	0	21	0
	F. vesca	1	0	0	1	0
	F. ananassa	423	1	56	366	1

correlations largely differed between the two species, as manifested in the lack of association between the metabolite coupling degrees. For instance, at a threshold value of 0.85, the metabolites with the largest coupling degrees in *E. coli* were: phenylanine, threonine, proline and lysine, while in *A. thaliana* they included: isoleucine, leucine, tyrosine and lysine (see Supplemental Table 8.2.4). It must be noted that these results and interpretations warrant caution since the metabolite profiles from *A. thaliana* were obtained from entire Arabidopsis rosette rather than from isolated chloroplast, which may bias the drawn conclusions. The analysis can be conducted with compartment-specific metabolic profiles once they become available.

5.3.3 SCA shows that domestication in wheat is associated with loss of regulatory couplings

Domestication of tetraploid wheats, *Triticum turgidum L.*, is an important evolutionary event for the human development. Emmer (*T. turgidum ssp. dicoccum*) was domesticated from wild emmer (*T. turgidum ssp. dicoccoides*) around 12,000 years ago [Nesbit and Samuel, 1998]. Free-threshing tetraploid wheats (*T. turgidum ssp. turgidum*) subsequently originated from emmer, followed by the selection of durum wheat (*T. turgidum ssp. turgidum convar. durum*). Therefore, it has been suggested that the evolution of tetraploid wheats consists of at least two steps: primary domestication, from wild emmer to emmer, and secondary domestication, from emmer to durum wheat [Gioia et al., 2015].

Since important domestication-associated traits (e.g. the increase in seed size, the loss of dormancy [Gepts and Papa, 2002]) often necessitate alteration of metabolic process, we asked if application of SCA to metabolic profiles can be used to quantify the effect of domestication with respect to loss or gain of regulatory couplings. To this end, we used recently analyzed data about the phenotypic variation of primary metabolites in the kernels from three *T. turgidum* populations that represent both

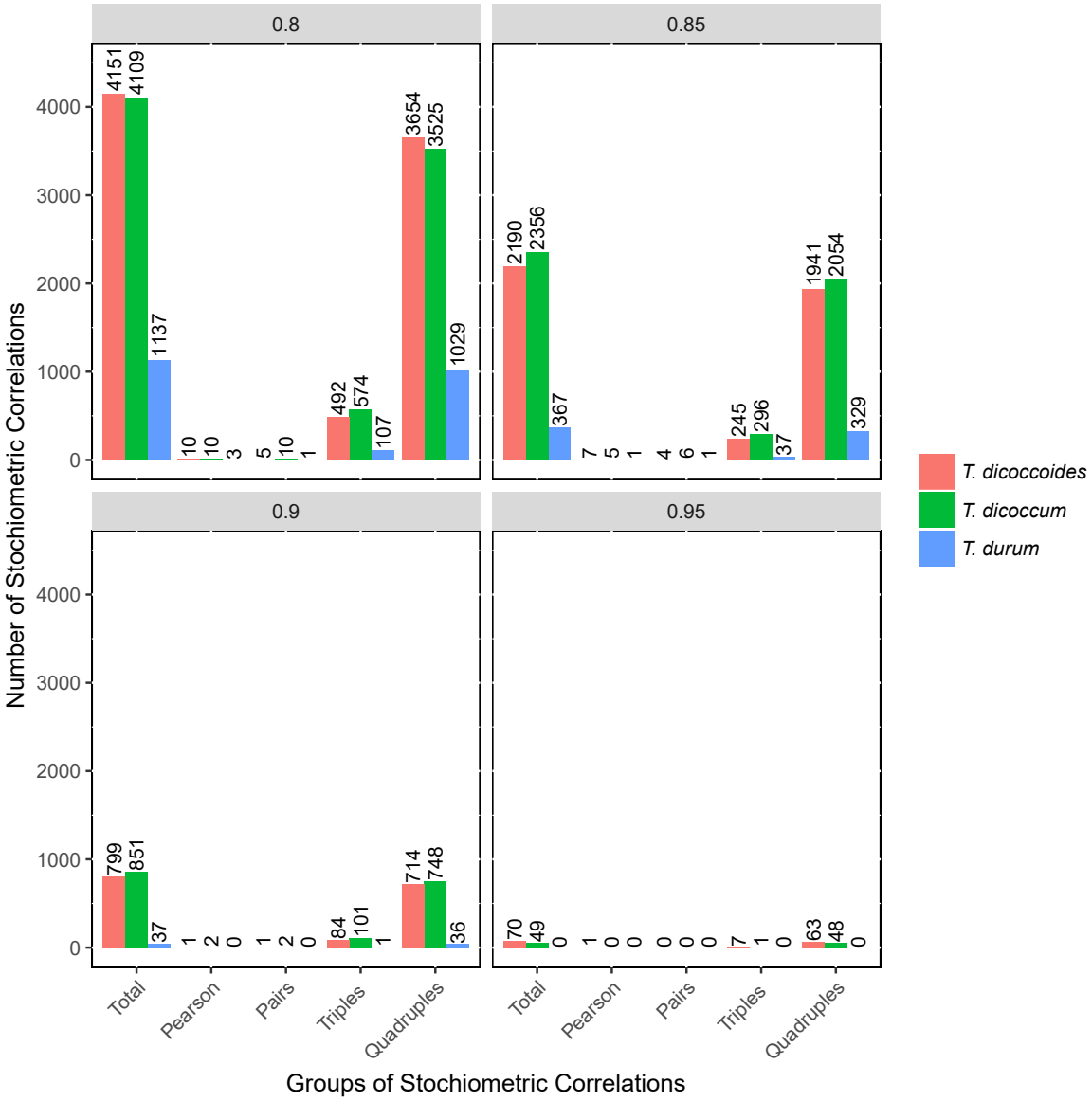

Figure 5.5: **Number of Stoichiometric Correlations of Wheat taxa**
Shown is the number of stoichiometric correlations at the four thresholds 0.8, 0.85, 0.9 and 0.95. The bars represent the the total number of stoichiometric correlations, pairs, triples and quadruples for all three wheat taxa. The exact values are shown above the bars.

the primary and secondary domestication [Beleggia et al., 2016]. Beleggia et al. [2016] determined that there were changes in content of specific metabolites, particularly amino acids and unsaturated fatty acids, associated with the primary and secondary domestication events. The resulting metabolic profiles of accessions within each taxon were also employed to construct Pearson correlation networks. Based on various properties of the correlation networks (e.g. shared correlations, centrality of metabolites) it was concluded that the difference between wild emmer and emmer were larger than the difference between wild emmer and durum wheat. In addition, it was found that durum wheat contained a larger number of significant correlations, followed by wild emmer and emmer. Therefore, surprisingly, the results from Pearson correlation analysis captured contrasting findings in comparison to the evolutionary distance between the three analyzed taxa.

We applied SCA to contents of 22 metabolites from four compound classes (i.e. amino acids, sugars, organic acids, and alcohols) within each population at four threshold values (see Section 5.2: Materials and Methods). These metabolites were selected based on their presence in every of the analyzed accessions to allow comparative analysis of the populations without the need of imputation as well as assumptions about the reasons for absence of detected metabolites. The number of stoichiometric correlations due to triples shared between emmer and wild emmer was the highest, followed by that between durum and wild emmer (at threshold of 0.8). At a threshold value of 0.85, both emmer and wild emmer had one overlapping triple with durum. However, at a threshold value of 0.8, durum wheat shared more stoichiometric correlations due to quadruples with wild emmer than emmer. At thresholds of 0.85 and 0.9, durum shared the same number of quadruples with emmer and wild emmer. In all cases, only stoichiometric correlations due to quadruples were shared between all three populations (e.g. stoichiometric correlations at threshold of 0.85, Supplemental Table 8.2.6). Overall, we observed more triples and quadruples in *T. dicoccum* and *T. dicoccoides* in comparison to *T. durum* (see Figure 5.5). The observation was supported by the quintiles of the correlation values shown in Supplemental Table 8.2.7.

This finding implied that the loss of traits due to domestication and increase in seed size were associated with an overall loss of reaction couplings reflected in the smaller number of stoichiometric correlations in durum wheat in comparison to (wild) emmer (Supplemental Tables 8.2.5, 8.2.6 and 8.2.7). The metabolites involved in the largest number of stoichiometric correlations above a threshold value of 0.85 in wild emmer included glycine, threonine, aspartate, serine and glutamate; in emmer, these metabolites included serine, leucine, threonine, and glutamate, while in durum wheat they consisted of fructose, glucose, glutamate, and asparagine (see Supplemental Table 8.2.7). Altogether, the application of SCA identified a shift in importance of regulatory role of sugars in comparison of organic and amino acids which is in line with the increase in seed size due to the need for more cell wall components.

To further validate our results from the three wheat taxa, we included data of six different tomato species into the analysis. We compared the domesticated *S. lycopersicum* (M82) to the group of the other five species, as their fruits drastically differ from those of M82 [Schauer et al., 2005]. However, it has to be noted that there is no clear linage from the undomesticated plants to the M82. Additionally, the combination of the different tomato species might result in an inclusion of additional noise. Nevertheless, it is a necessary step to have the needed amount of replicates per metabolite. Overall, the tomato data set contains 43 metabolites common to the analyzed species. In line with the results of wheat, we observed fewer stoichiometric correlations for M82 than for the undomesticated wild-type tomato (Table 5.2 and Supplemental Table 8.2.8). The exception is the threshold of 0.8; in this case, the

M82 has roughly 4000 more pairs, triples and quadruples than the wild-type tomato. At a threshold value of 0.85 the M82 has still around 300 stoichiometric correlations more than the wild-type species. With increasing threshold, however, the number of significant stoichiometric correlations decreases in M82 more than in the wild type. This finding was reflected in the different quintiles of the correlation values for the domesticated and wild tomato species (Supplemental Table 8.2.2). The metabolites with the largest number of stoichiometric correlations above a value of 0.9 in wild-type tomato are erythritol, cysteine, succinic acid and beta-alanine, while in M82, they include: leucine, putrescine, dehydroascorbic and sucrose (Supplemental Table 8.2.4).

A very similar scenario was considered with the strawberry accessions *F. vesca* (wild) and *F. ananassa* (domesticated and commercially available) without direct domestication lineage between the two species. In contrast to our observations in wheat and tomato, the domesticated strawberry exhibits a higher number of stoichiometric correlations above all thresholds (Table 5.2 and Supplemental Table 8.2.9). The reason for this finding may lay in the different ploidy of the investigated organisms, namely, *F. ananassa* is an octaploid organism, whereas *F. vesca* is diploid with a rather small genome. The application of SCA to metabolomics data from domestication implies a new principle which underlies this agronomically and evolutionary important process; namely, optimizing a given trait could be accomplished by breaking the existing regulatory mechanisms, reflected in the coupling of the biochemical reaction rates, which in turn provides a greater space of possibilities on which selection can operate.

5.4 Conclusion

Here we proposed a constrained extension to the concept of maximal correlation, based on the concept of reaction rate coupling in networks of metabolic reactions. The concept of reaction couplings forms the core of the stoichiometric correlation analysis. The constraints in the maximal correlation are due to the values which the linear combinations of log-transformed metabolic profiles are allowed to take. SCA facilitates the comparison of data sets on the same metabolites between two scenarios with the idea of comparing and contrasting the degree of coupling. By determining the stoichiometric correlations of metabolic profiles from the TCA cycle and amino acid synthesis, we showed that *E. coli* stringent response is differently (and less strongly) controlled than that of *A. thaliana*. Therefore, while the enzymes underlying the stringent response are preserved in these two model organisms, their integration in the metabolic networks may have evolved different regulatory action. In addition, SCA can be used to investigate the differences between wild and domesticated species, and to determine if the difference can be ascribed to alterations in metabolic couplings brought about by various regulatory mechanisms. Based on this idea, we demonstrate that stoichiometric correlations from metabolic profiles from natural variation in wild and domesticated species indicate that domestication is associated with loss of regulatory control. Therefore, our findings provide the basis for future flux-oriented studies towards mechanistic understanding of this important evolutionary process.

6 Data reduction approaches for dissecting transcriptional effects on metabolism

Publication: Front. Plant Sci. (under review)

Authors: Kevin Schwahn[1,2], Zoran Nikoloski[1,2,3]*

Affiliations: [1]Systems Biology and Mathematical Modeling Group,
Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, Potsdam-Golm, Germany

[2]Bioinformatics Group, Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24-25, Potsdam-Golm, Germany

[3]Center of Plant Systems Biology and Biotechnology (CPSBB), Plovdiv, Bulgaria

Contact: *nikoloski@mpimp-golm.mpg.de

Abstract

The availability of high-throughput data from transcriptomics and metabolomics technologies provides the opportunity to characterize the transcriptional effects on metabolism. Here we propose and evaluate two computational approaches rooted in data reduction techniques to identify and categorize transcriptional effects on metabolism by combining data on gene expression and metabolite levels. The approaches determine the partial correlation between two metabolite data profiles upon control of given principal components extracted from transcriptomics data profiles. Therefore, they allow us to investigate both data types with all features simultaneously without doing preselection of genes. The proposed approaches allow us to categorize the relation between pairs of metabolites as being under transcriptional or post-transcriptional regulation. The resulting classification is compared to existing literature and accumulated evidence about regulatory mechanisms of reactions and pathways in the cases of *Escherichia coli*, *Saccharomycies cerevisiae*, and *Arabidopsis thaliana*.

6.1 Introduction

Metabolism is the integrated output of transcription, post-transcriptional processes, translation and post-translational processes, and reflects the environment and changes in the availability of nutrients [Stitt, 2013; Johnson et al., 2016]. The combined outcome of the aforementioned processes is the metabolic state of the system, observed in its metabolite and enzyme levels as well as the resulting reaction rates. The rates of metabolic reactions, however, are difficult to monitor and require involved computational integration of data and models [Tang et al., 2009; Sims et al., 2013]. With advances in high-throughput techniques for monitoring of both metabolite and gene expression levels, the biological community is faced with the challenge of evaluating and integrating the obtained large-scale data to address several pressing questions: (1) which parts of metabolism are under regulation from transcriptional and downstream processes? [Less and Galili, 2008; Moxley et al., 2009; Haverkorn van Rijsewijk et al., 2011], (2) how metabolism feeds back to transcription to coordinate the systemic functions? [Pego et al., 2000; Ladurner, 2006; Kresnowati et al., 2006; Lu et al., 2007; Kochanowski et al., 2017], and (3) how and why are the changes at different cellular layers, like transcription and metabolism, suppressed or propagated to other layers? [Price et al., 2004; Ledezma-Tejeida et al., 2017; Gonçalves et al., 2017]. In this context, we ask to what extent purely statistical techniques can be used to investigate whether data from metabolomics platforms in combination with transcriptomics data corroborate existing findings or yield new insights into transcriptional control of metabolism.

Microarray and RNA-sequencing techniques can measure several thousand genes from multiple conditions and time points simultaneously [Meyers et al., 2004; Jain, 2012]. In contrast, metabolomics platforms provide measurements for only a fraction of the metabolon, including all metabolites in a given system [Fernie et al., 2004]. Despite the growing number of publically available data sets, the case in which both data types are available from the same experiments is limited to only few observations (i.e. experiments, time points, replicates). Therefore, any method which is used to jointly investigate transcriptomic and metabolomic data faces the problem of high dimensionality of both data types and difference in the number of components measured. As a result, various multivariate statistics approaches have been evaluated to facilitate the analysis of transcriptomic and metabolomic data from the same experiments.

Whatever the multivariate statistical approach used, its aim is to identify an association between one or more genes and one or more metabolites. As a result, we can classify the methods into those which establish an association between (1) single gene and single metabolite, (2) multiple genes and a single metabolite, (3) single gene and multiple metabolites, and (4) multiple genes and multiple metabolites. The first set of approaches is the simplest and aims at identifying the association for a pair of a gene and a metabolite [Tohge et al., 2015; Cavill et al., 2016] by applying different similarity measures, such as: Pearson and Spearman correlation [Urbanczyk-Wochniak et al., 2003; Gibon et al., 2006; Hannah et al., 2010] or time-shifted correlations, in case when time-series data are analyzed [Walther et al., 2010; Takahashi et al., 2011]. A general observation is that there is a high number of correlations between transcripts and metabolites, rendering it challenging to determine molecular/cellular mechanisms, and that one metabolite correlates to multiple transcripts, likely due to pleiotropic effects [Urbanczyk-Wochniak et al., 2003; Hannah et al., 2010]. Further, the type of observed correlation (positive or negative) highly depends on the experimental condition. Along these lines, the resulting associations have also been analyzed with methods from network analysis [Bradley et al., 2008; Redestig and Costa, 2011]. Approaches based on correlation networks have been employed for annotation of gene function given information about the compound class and structure of the metabolite [Tohge and Fernie, 2010, 2012].

In the case where association between multiple genes and one metabolite is to be identified one can use several approaches. For instance, classical regression techniques can be readily employed, with additional regularization to address the issue of high-dimensionality of the data [Auslander et al., 2016]. On the other hand, dimension reduction techniques coupled with network analysis can be used to identify such associations: For instance, Inouye et al. [2010] identified modules of coexpressed genes on which they perform principal component analysis (PCA). The association (per Spearman correlation) between the first principal component and a given metabolic data profile is used as a means to determine which modules have influence on the metabolite level. While regression-based analysis is unbiased, in the sense that it can include all measured genes, it requires large data sets for estimation of the model coefficients. On the other hand, the identified modules based on correlation may be change if new data are analyzed, indicating bias in the identified associations. In principal, by symmetry, these approaches can be used to identify and study associations between multiple metabolites and a single gene [Kochanowski et al., 2017].

The most involved cases are those where associations are to be established between multiple genes and multiple metabolites. In this case, there have been several approaches developed and used in the joint analysis of transcriptomics and metabolomics data sets: Partial Least Squares (PLS) and its extensions [Bylesjö et al., 2007] and canonical correlation analysis (CCA) [Jozefczuk et al., 2010]. PLS aims to find the relation between two matrices X and Y by estimating the direction in X that explains most of the variance in Y [Boulesteix and Strimmer, 2007]. Due to its multivariate nature, PLS regression is difficult to interpret. The orthogonal projections to latent structures (OPLS) was designed to improve the interpretation of the regression. The approach allows to remove variation form X, which is uncorrelated (orthogonal) to Y. The advantage compared to PLS is twofold: first, the orthogonal part of X can be separately investigated and secondly and more important the removal of uncorrelated variation increases the interpretation [Trygg and Wold, 2002; el Bouhaddani et al., 2016]. Given two data sets X and Y, CCA finds the canonical variates, $U = a'X$ and $V = b'Y$, so that the correlation between U and V is maximized. The advantage of CCA is, that it is invariant with respect to transformation of the variables. However, the calculation of the CCA requires the inverse of XX^T which is challenging

when the number of transcripts or metabolites exceeds the number of observations (as is the case for most biological studies). A solution to this dimensionality problem is to focus on a subset of the data, so that the number of transcripts (metabolites) is smaller than the number of observations, which may introduce bias in the analysis [Jozefczuk et al., 2010].

While the four classes of approaches can determine association between a subset of genes and a subset of metabolites, they cannot be used to determine if the relation between two metabolites is under transcriptional or post-transcriptional control. This question goes beyond the analysis of the effects of transcripts on the level of metabolites, but rather on the coordination between metabolite levels. Addressing this issue will shed light on the transcriptional control of metabolic coordination. To this end, we propose two approaches rooted in a combination of partial correlation and dimension reduction techniques. We tested our proposed approaches with data sets from *Escherichia coli*, *Saccharomycies cerevisiae*, and *Arabidopsis thaliana* to identify metabolite pairs which are associated either by transcriptional or post-transcriptional regulatory effects. Our proposed approach might be used for biotechnology studies, where it can suggest metabolites whose relationship is under transcriptional regulation and is therefore easier to manipulate through genetic engineering.

6.2 Materials and Methods

6.2.1 Data used in the study

The data used in this study were downloaded from the supplementary of Jozefczuk et al. [2010] and contain metabolomic and transcriptomic data from *E. coli* under different conditions (cold, heat, change from glucose to lactose and oxidative stress) as well as control treatment. The metabolomic data were generated by gas chromatography-mass spectroscopy (GC-MS) and contain 192 metabolites. Transcript data were measured with a microarray based technique and 4440 transcripts were detected. In total 82 common data points were used for the analysis. Additionally, *A. thaliana* data from the study of Caldana et al. [2011] were used. The metabolomic data were generated by GC-MS and consists of 92 metabolites measured under the following conditions: 21° C at 75 $\mu E m^{-2} s e c^{-1}$, 150$\mu E m^{-2} s e c^{-1}$ light intensity and darkness, 4° C at 85 $\mu E m^{-2} s e c^{-1}$ light intensity and darkness, 32° C at 150 $\mu E m^{-2} s e c^{-1}$ and darkness. Therefore, the analyzed data set consists of metabolic time series covering 20 time points and gathered under seven different light and temperature combinations. Further, a data set from *S. cerevisiae* containing metabolomic and transcriptomic data from three different growth conditions, nitrogen upshift (shift from proline two glutamine), nitrogen downshift (shift from glutamine two proline) and Rapamycine treatment, was included. The data set contains 256 metabolites measured with FIA-QTOF-MS and 5716 transcripts measured with Affymetrix chips [Oliveira et al., 2015]. As the dimensions of the two data sets, i.e., transcripts and metabolites, need to agree, only matching time points per experiment were taken into account. The complete list of experiments and the appropriate time points is provided in Supplemental Table 8.3.1. In total 41 data points were used per metabolite and transcript.

6.2.2 PCA and partial correlation

PCA is a statistical procedure that uses an orthogonal linear transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called

PCs. The PCs are ordered according to the variance they explain [Wold et al., 1987]. Partial correlation measures the relationship (correlation) of two variables while controlling for a third or more variables. When using a single controlling factor, one calculates the first order partial correlation. If the number of controlling factors is higher, their information is recursively removed and the second or higher order partial correlation is determined. The zero order partial correlation is the same as the Pearson correlation. The expression for recursive calculation of partial correlation between the variables X and Y given a set of controlling variables in V is given by

$$r_{XY.V} = \frac{r_{XY.V/Z} - r_{XZ.V/Z}\, r_{YZ.V/Z}}{\sqrt{1-r^2_{XZ.V/Z}}\sqrt{1-r^2_{YZ.V/Z}}}. \tag{6.1}$$

where $Z \in V$.

6.2.3 Combination of PCA and partial correlations to investigate influence of transcripts on metabolites

The combination of partial correlation and PCA allows the calculation of the two approaches *Transcriptional dependent Partial Correlation* and *Post-transcriptional dependent Partial Correlation.* We compute the first p PCs of the transcript data and use them as controlling variables for the partial correlation for each combination of metabolites. The number of PCs to choose was investigated based on the Broken-Stick model [Jackson, 1993], Kaiser-Guttman criteria [Yeomans and Golder, 1982] and the Horn's parallel analysis (PA) [Horn, 1965; Dinno, 2009]. For the Broken-Stick model a distribution is calculated $\lambda_i = \sum_{k=i}^{p} \frac{1}{k}$, where p is the number of variables and λ_i the eigenvalue of the i^{th} component [Jackson, 1993]. In the Kaiser-Guttman approach, the PCs with an eigenvalue above the mean of the eigenvalues are regarded as significant [Yeomans and Golder, 1982]. We performed Horn's parallel analysis by randomizing the transcript data and calculating the eigenvalues for the randomized data. A PC is identified as significant if its eigenvalue is larger than a chosen percentile of the distribution of eigenvalues of that component. We performed 1000 randomization and regarded a PC as significant, if it exceeds the 99 percentile of the distribution of eigenvalues. We then compute significant differences of Pearson correlation and in-significant partial correlation pairs after removing the first p representative PCs from the transcriptomic data, yielding *Transcriptional dependent Partial Correlation.* This gives transcriptionally regulated pairs of metabolites. In contrast we can use the same first p representative PCs to calculate the *Post-transcriptional dependent Partial Correlation*, using the significant differences of Pearson correlation and significant partial correlations.

6.2.4 Calculating significant differences with permutation test

Testing for significant interactions of metabolites was performed by permutating the transcript and metabolite data component-wise. Calculations based on the two approaches are repeated for each of the 5000 permutations. For all approaches we adjusted for multiple hypothesis testing, using Benjamini-Hochberg with a significance level $\alpha = 0.01$.

6.2.5 Algorithm Implementation

All analysis were performed in R [R-Core-Team, 2013] using the default functions and the corr.test() function of the psych package. For the recursive calculation of the partial correlation the pcor.rec() function was used, downloaded at `http://www.yilab.gatech.edu/pcor.R`. Evaluation of the permuted data to determine significance was implemented as stand-alone function. The estimation of the Kaiser-Guttman and the Broken-Stick model were done based on the provided function in the supplemental material of Borcard et al. [2011].

6.3 Results

6.3.1 Two novel methods for categorization of metabolite pairs based on transcriptional effects

In this study we developed two approaches that allow the simultaneous investigation of transcriptomic and metabolomic data from the same experimental setup (see Figure 6.3.1). The novelty of the proposed approaches lies in the way the transcriptomic data are used to partial out (remove) the effect of the transcription layer from the metabolitc layer. Partial correlation has been used to investigate large-scale data sets from different *omics* technologies [de la Fuente et al., 2004; Ursem et al., 2008; Veiga et al., 2007; Wu et al., 2013]. Partial correlation quantifies the association between two variables, while controlling for the influence of another set of variables. Therefore, it has been helpful in identifying non-spurrious associations [Baba et al., 2004]. However, as higher order partial correlations are calculated iteratively, the calculation quickly becomes unfeasible with large transcriptomic or metabolomic data sets.

Our first approach, termed *Transcriptional dependent Partial Correlation* (TPC), aims at identifying pairs of metabolites whose association is related to transcriptional regulation. The approach is composed of four steps: (1) We calculate the first p principal components (PCs) of the trancriptomic data, (2) we determine all metabolite pairs having a significant Pearson correlation coefficient, (3) we determine all metabolite pairs having non-significant partial correlation upon removal of the controlling variables, i.e., the p PCs from step (1), above, and (4) for the pairs of metabolites in the sets obtained from (2) and (3), we select those that show a significant difference between their Pearson correlation and partial correlation values. The reason for such construction of the approach is the following: if the removal of the significant PCs leads to a non-significant partial correlation between two metabolites, their association was due to transcriptional regulation. As the significant PCs capture most of the transcriptional effects, by finding the partial correlations we remove most of the transcriptional influence on the association between the two metabolites. The statistical tests ensure that the consideration of the significant PCs indeed break the significant association between metabolites and that the difference between the values is significant. To determine statistical significance we rely on permutation tests (see Section 6.2: Materials and Methods).

The second approach, termed *Post-transcriptional dependent Partial Correlation* (PPC), follows a similar methodology. Again, the significant PCs from the transcriptomics data set are used as control variables in the partial correlation analysis for pairs of metabolites. In contrast to the TPC approach, we select those pairs of metabolites that are significantly associated upon removal of the significant PCs.

Similar to TPC, we select those pairs of metabolites whose partial correlations show significant difference from the values of the respective Pearson correlation coefficient. The approach is based on the premise that if correlation remains upon removal of the transcriptional effect, the observed association is due to post-transcriptional regulation of the two metabolites. The significant difference to the observed Pearson correlation is employed to ensure that the observed partial correlation is due to post-transcriptional effects.

As both approaches relies on the estimation of principal components from the transcriptiomic data, the question arises, how many should be used for the analysis? More components will increase the computation time, while a to small number of PCs will not integrate sufficient transcriptomic information into the analysis. Multiple approaches have been reported to estimate the significant PCs. We employed the Kaiser-Guttman criteria [Yeomans and Golder, 1982], the Broken-Stick model [Jackson, 1993] and Horn's parallel analysis (PA) [Horn, 1965; Dinno, 2009] (see section 6.2: Materials and Methods). Overall, we used our TPC and PPC approach on three different data sets, namely from *E. coli, S. cerevisiae* and *A. thaliana* (see Section 6.2: Materials and Methods). To this end, we investigated the number of significant PCs for each of the available data sets. The Kaiser-Guttman approach suggested the use of three PCs in each of the three data sets, whereas the Broken-Stick model suggested the usage of only one PC for *E. coli* and *A. thaliana* and two for *S. cerevisiae.* The PA approach confirmed the one PC for *E. coli* and two for *S. cerevisiae.* However, the approach estimated two significant PCs for the *A. thaliana* data set. Overall, we found between one and three significant PCs, depending on the approach and data set (see Supplemental Figure 8.3.1). Therefore, we decided to use three PCs as a good compromise between the variance of the transcript data explained and the running time of the algorithm.

6.3.2 Transcriptional and post-transcriptional control of metabolite associations in *E. coli*

In this section, we applied our approaches with a transcriptomics and metabolomics data set from *E. coli* (see Section 6.2: Materials and Methods), containing the levels of 192 metabolites and 4400 genes over five conditions. Employing the TPC resulted in 87 metabolite pairs under transcriptional control (Supplemental Table 8.3.2), whereas 739 metabolite pairs were found to be under post-transcriptional control with the PPC approach (Supplemental Table 8.3.3). As a first control, we did not identify an overlap between the pairs of metabolites detected with the two approaches.

In a first general investigation we found no change in the sign between Pearson and partial correlation. However, we investigated, if the absolute value of the correlation increased or decreased upon performing the partial correlation (Figure 6.2). Most of the significant correlations had a lower value when using partial correlation, in comparison to the Pearson correlation. More than 80% of the positive correlations found with the PPC approach decreased and around 60% of the positive correlations from the TPC approach. However, the overall observed differences between the values of Pearson and partial correlation were between 0.005 and 0.03. Although, we did observe a change in the correlation with our approach, the magnitude is small.

In the following, we focused on the analyses of annotated metabolite pairs to allow for a comparison to the results previously reported in the literature. Out of the 87 metabolite pairs from the TPC approach 19 metabolite pairs (Supplemental Table 8.3.4) were unambiguously identified, whereas 132 of the 630

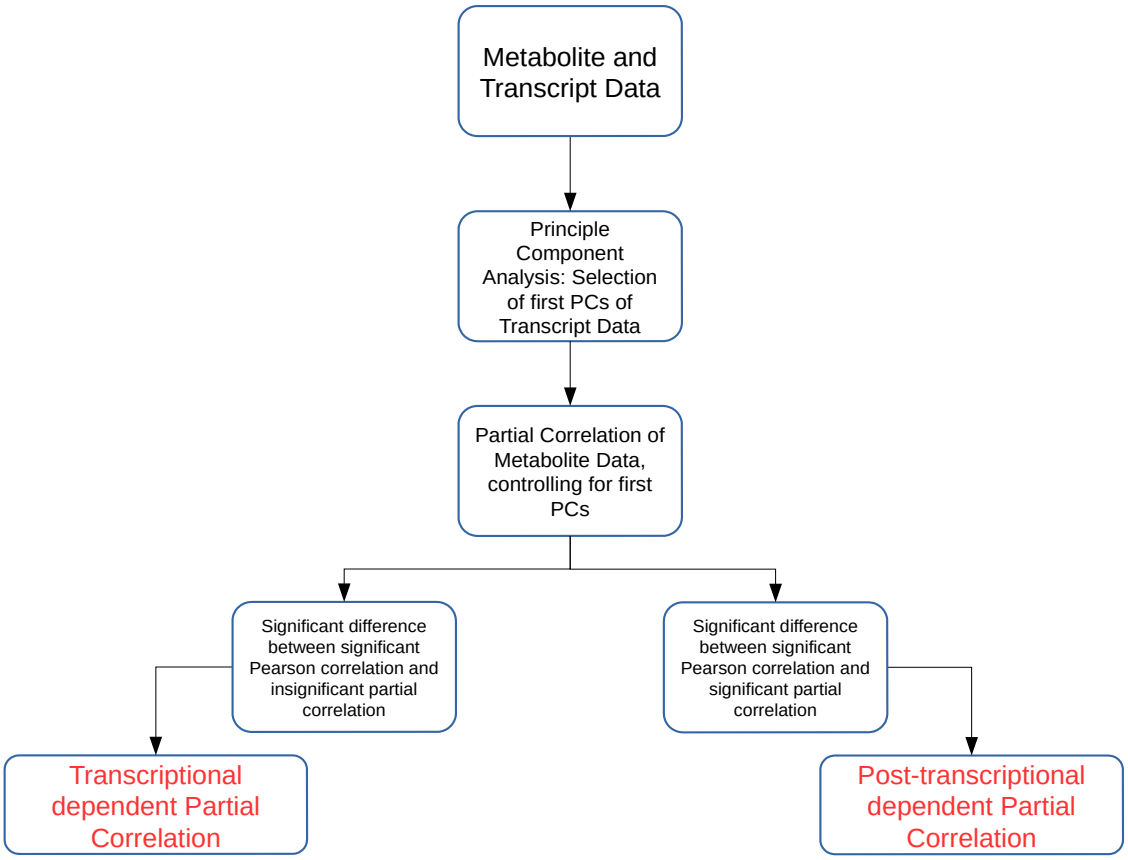

Figure 6.1: **Schematic overview of the two approaches introduced in the study.**
The approaches *Transcriptional dependent Partial correlation* (TPC) and *Post-transcriptional dependent Partial correlation* (PPC) use the first p PCs of the transcriptomic data as control variables in the partial correlation calculation.

pairs of the PPC approach (Supplemental Table 8.3.5) were unambiguously identified. For instance, phosphate and maltose showed Pearson correlation of -0.26 and a partial correlation of -0.25. Both metabolites are part of the phosphoenolpyruvate dependent phosphotransferase system (PTS). The system consists of three enzymes performing the phosphate transport from PEP onto a carbohydrate. Maltose is one of the acceptors and belongs to the Glucose-class within the PTS. The three enzymes in the PTS are EI, Hpr and EII, which are encoded on the pts-operon, which itself is transcriptional regulated and induced through glucose. We therefore found a metabolite pair participating in a fully transcriptional regulated pathway [Postma et al., 1993; Tchieu et al., 2001]. The weak negative correlation is explained by the fact that PEP acts a the main phsophate donor and we therefore capture not the complete active level of phosphates in this pathway. The negative correlation is explained by the reversibility of the system. While the sugar, here maltose, is only involved in one of the reactions within the PTS, the phosphate can be transferred between the three proteins [Deutscher et al., 2006]. Therefore, an increase of maltose will have a delayed effect on the phosphate pool.

Further, we investigated the literature regarding GABA and L-ornithine showing a Pearson correlation of -0.30 and a partial correlation of -0.28. The negative correlation is related to the fact that the two metabolites are competing substrates for the same enzyme. The processing of one metabolite by the enzyme leads to an accumulation of the other substrate, which was shown in simulation studies [Schäuble et al., 2013]. GABA and L-ornithine are connected via the enzyme 4-aminobutyrate aminotransferase, since it can use GABA and N-acetyl-L-ornithine as substrates [Lal et al., 2014] and N-acetyl-L-ornithine can be transformed into L-ornithine by one additional reaction. The enzyme, 4-aminobutyrate aminotransferase, is encoded by the gene gabT [Kurihara et al., 2010] and is activated by the regulatory protein cAMP receptor protein (CRP) [Metzner et al., 2004]. CRP regulates gabTs activity mostly under stress conditions, more precisely at starvation. Further, regulatory mechanisms that influence the expression of gabT are the sigma factors sigmaS and sigma38 which are encoded by the gene RpoS [Joloba et al., 2004]. Therefore, it is expected that the association between GABA and L-ornithine is transcriptionally regulated by CRP and RpoS.

Finally, we investigated the identified pair of 3PGA and aspartate. For these metabolites, the observed Pearson correlation was 0.28 and the partial correlation was 0.27. Jozefczuk et al. [2010] reported that in general 3PGA decreases under stress conditions, while only under cold stresses aspartate levels increase. The weak positive correlation is likely due to the fact that we investigated the correlation over multiple conditions [Bradley et al., 2008]. Aspartate can be synthesized out of oxaloacetate, which additionally stands in exchange with 3PGA through PEP within the glycolysis. We therefore are able to identify two metabolites from the same pathway separated by two reactions, taking part in the glycolysis and the TCA cycle. Both pathways are partially regulated on the transcriptional level. For instance, the transcription factor Cra is involved in feedback and feed-forward regulation within these pathways [Shimizu, 2013]. It is activating the transcription of the gene coding for the enzyme isocitrate dehydrogenase, which is an essential step for the transformation from citrate to all further downstream metabolites in the TCA cycle [Prost et al., 1999]. Overall, the regulatory process will influence the pair of 3PGA and aspartate.

To come to a general conclusion, we investigated the available literature involving the metabolite pairs identified by the PPC approach. In comparison to the TPC approach, we frequently found amino acids within the pairs of the PPC approach (Supplemental Table 8.3.3). As amino acids are regulated through feedback inhibition by their loaded tRNAs [Sanchez and Demain, 2008], our approach captured the post-transcriptional regulation. For further validation, we investigated the literature regarding the

two pairs of PEP-valine (Pearson correlation of -0.35 and partial correlation of -0.37) and PEP-leucine (Pearson correlation of -0.37 and partial correlation of -0.39). The negative correlation of PEP and the amino acids leucine and valine were previously reported in Szymanski et al. [2009] under stress conditions, which are comparable to the experimental conditions from our data set. The PEP generating enzyme, the pyruvate kinase, is inhibited by fructose 1,6-bisphosphate and structural similar metabolites [Speranza et al., 1990]. The synthesis of PEP is therefore under strong post-transcriptional regulation. Additional, the both mentioned amino acids are produced from pyruvate. Pyruvate is altered into PEP by a reversible reaction linking it further to post-transcriptional regulation. Further, valine and leucine share part of their synthesizing pathways. Valine is involved in a feedback inhibition of the enzyme acetohydroxy acid synthase and inhibits the leucine and the isoleucine synthesis as well. Furthermore, leucine inhibits its own producing enzymes (a-isopropylmalate synthase) regulating the group of amino acids coming from pyruvate. All three metabolites of the pairs are under post-transcriptional regulation. In addition, we found metabolites belonging to the TCA cycle and related reactions. Among these metabolites are malate, fumarate, PEP, 3PGA and GABA. Out of these malate and PEP (Pearson correlation of -0.29 and partial correlation of -0.31) were previously reported to be negatively correlated [Szymanski et al., 2009]. PEP level increases under stress, while malate and precursors decrease. In contrast, the pair of 3PGA and GABA are positively correlated (Pearson correlation of 0.30 and partial correlation of 0.29). 3PGA level were reported to decreases under stress [Jozefczuk et al., 2010], while Szymanski et al. [2009] reported that amino acids decreased under stress conditions which will affect GABA as well. The prevailing regulatory mechanism in the TCA cycle are product inhibition, substrate availability and competitive feedback inhibition. The citrate synthase is inhibited by citrate, further Succinyl-CoA is a competitor with acetyl-CoA for the citrate synthase as well. The first example is a product inhibition, whereas the second example is competitive feedback inhibition. Further, the isocitrate dehydrogenase is regulated by phosphorylation in *E. coli*. After phosphorylation the enzymes becomes inactive. Therefore, the TCA cycle is highly regulated on the post-transcriptional level [Voet and Voet, 2011]. We can therefore confirm that malate, PEP, 3PGA and GABA are under post-transcriptional regulation.

Our approach allows to distinguish between metabolite pairs with associations controlled at transcriptional or post-transcriptional level. Therefore, we extended our analysis to data sets of *S. cerevisiae* and *A. thaliana*, aiming to reproduce the classification of metabolite pairs into transcriptional and post-transcriptional associated at higher organism.

6.3.3 Prevailing regulatory effects in *S. cerevisiae* - comparison with published results

So far we were able to identify the prevailing regulatory mechanism between identified pairs of metabolites. However, our comparison focused on a broad literature comparison, but did not compared our approach directly with a comparable method capable of integrating transcriptomic and metabolomic data into a combined analysis. Therefore, we chose to complement our study with a comparison with the results obtained in Oliveira et al. [2015]. The study investigated the regulatory effect occurring during a nitrogen supply shift (upshift and downshift) as well as the treatment with Rapamycin in *S. cerevisiae*. Metabolite and transcript data were measured at up to 19 time points for the metabolite data and up to eight time points for the transcript data for each of the three conditions. The overlapping eight time points are therefore ideal for our proposed method. In the origi-

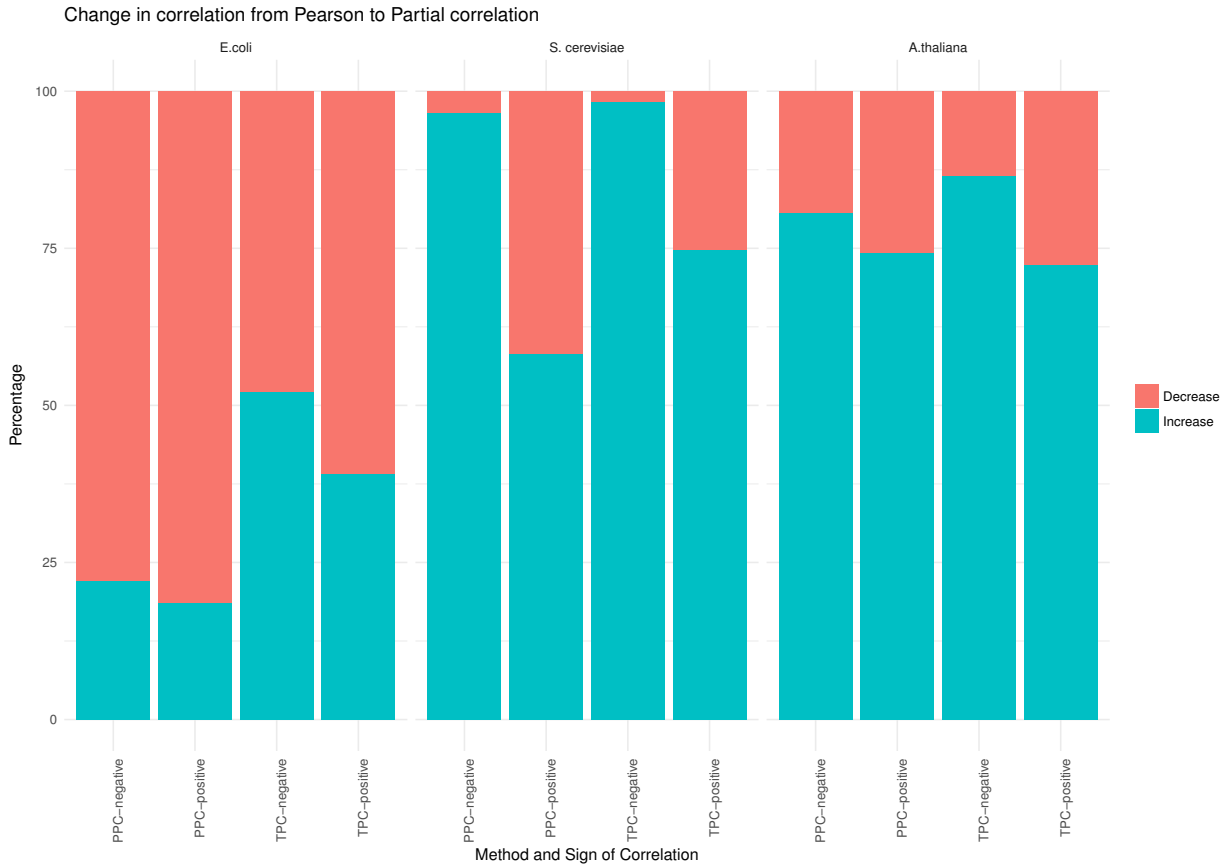

Figure 6.2: **Changes from Pearson to partial correlation.**
Changes from Pearson to partial correlation for all three organism (*E. coli, S. cerevisiae* and *A. thaliana*) and for TPC - positve correlation, TPC - negative correlations, PPC - positive correlations and PPC - negative correlations. The blue portion of the bar represents the percentage of significant correlations whose absolute value increased from Pearson to partial correlation. The red portion of the bar represents the percentage of significant correlations whose absolute value decreased from Pearson to partial correlation.

nal study, the authors used Bayesian inference to assign each metabolite to one of the four network motifs "unrelated" (no regulation related to TORC1), "downstream" (metabolites post-translational regulated downstream of TORC1), "upstream" (transcriptional regulation by metabolites upstream of TORC1) and "parallel" (transcriptional regulation by metabolites parallel of TORC1). The assignment of metabolites into on of the four categories is done by evaluating the dynamic dependence of metabolite and transcript pairs over time and the association of each metabolite with a specific set of genes regulated by TORC1, called "representation of TOR genes". Both features are combined in a Bayesian inference framework approach to calculate the probability for each metabolite to belong to one of the four motifs. If the probability is above 50%, the metabolite is assigned to that particular motif. Eight metabolites were assigned to the downstream motif, eight metabolites to the upstream motif and eight metabolites to the parallel motif.

We used their provided data and applied our approaches resulting in 1221 unique pairs with the TPC approach (Supplemental Table 8.3.6) and 4239 unique pairs with the PPC approach (Supplemental Table 8.3.7). We compared the "downstream" assigned metabolites with our PPC approach, whereas the motifs "parallel" and "upstream" both relate to transcriptional regulation and were compared to our TPC approach. Similar to the results of *E. coli*, we observed no change in the sign of correlation between Pearson and partial correlation. In addition, we report the number of significant correlations above certain thresholds in Table 6.1. We observed higher significant correlations with the PPC approach for positive correlations, as well as for negative correlations, in comparison to the TPC approach. We found correlation above 0.9 with the PPC approach, whereas the correlations of the TPC did not exceed 0.65.

Within the eight metabolites assigned in the "downstream" motif, we found 10 metabolite pairs with the PPC approach (see Table 6.2). Only trehalose-6phosphate and tetracosanoate are not part of any pair. Each of the remaining metabolites was part of at least two and up to four pairs. We found 16 metabolites of the "upstream" and "parallel" motifs, and we identified 11 pairs between these metabolites with the TPC approach (see Table 6.3). Only two pairs were found within the "parallel" group, the remaining nine pairs were between the groups "upstream" and "parallel".

Overall, our approaches were able to categorize all investigated metabolites into transcriptionally or post-transcriptionally associated. In contrast, in the study of Oliveira et al. [2015] the majority of metabolites were assigned to the "unrelated" motif or none. The main reason is that their study focuses on TORC1 dependent regulation, while our approaches integrate all regulatory effects given the available data sets. We can therefore give a comprehensive overview of the regulatory mechanism affecting the associations in which each metabolite is involved.

6.3.4 Transcriptional control of metabolite associations in *A. thaliana*

We also investigated a data set from the model plant *A. thaliana* containing the levels of 92 metabolites and 15089 genes over 7 conditions (see Section 6.2: Materials and Methods). Within this data set, we found 295 transcriptional associated metabolite pairs with the TPC approach (Supplemental Table 8.3.8). The PPC approach yield in total 1534 metabolite pairs under post-transcriptional control (Supplemental Table 8.3.9). Similar to the results of the two previous investigated data sets, we did not observe a change in the sign of the correlation from Pearson to partial correlation. In contrast to *E. coli*, we observed metabolite pairs with a higher absolute partial correlation value than Pearson correlation value (Figure 6.2). We found more than 72% of the positively correlated metabolite pairs

Table 6.1: Number of significant correlations above certain thresholds for the TPC and PPC approach for the data of Oliveira et al. [2015].

Threshold	Number of significant TPC correlation	Number of significant PPC correlation
> 0.9	0	4
> 0.8	0	151
> 0.7	0	567
>0.6	2	1170
>0.5	43	2152
>0.4	456	3260
< -0.4	49	495
< -0.5	1	125
< -0.6	0	16
< -0.7	0	0

Table 6.2: Metabolite pairs found within the downstream motif of the approach by Oliveira et al. [2015] and our PPC approach

Metabolite 1	Metabolite 2	Pearson correlation	Partial correlation
Pyrroline-3H-5C	Adenosine	0.484	0.433
Pyrroline-3H-5C	dGuanosine	0.483	0.433
Pyrroline-3H-5C	IMP	0.759	0.736
Indole-3-acetate	Adenosine	0.584	0.538
Indole-3-acetate	dGuanosine	0.584	0.538
Indole-3-acetate	IMP	0.616	0.571
Adenosine	IMP	0.744	0.708
Adenosine	L-Aspartate	-0.471	-0.418
dGuanosine	L-Aspartate	-0.471	-0.418
dGuanosine	IMP	0.744	0.701

Table 6.3: Metabolite pairs found within the downstream motif of the approach by Oliveira et al. [2015] and our TPC approach

Metabolite 1	Metabolite 2	Pearson correlation	Partial correlation
NAD	AICAR	0.468	0.508
Thiamin triphosphate	AICAR	0.468	0.392
Thiamin triphosphate	L-Leucine	0.367	0.397
Thiamin triphosphate	5-L-Glutamyl-L-alanine	0.452	0.398
Ornithine	Dihydroxyacetone	-0.415	-0-377
Ornithine	Glyceraldehyde	-0.415	-0.377
Ornithine	D-Lactate	-0.415	-0.377
Ornithine	Imidazole glycerol-P	-0.374	-0.289
L-Leucine	AICAR	0.434	0.462
GABA	Glyceraldehyde	-0.384	-0.348
GABA	Glutamine	-0.388	-0.319

identified with TPC, more than 74% of the positively correlated metabolite pairs identified with PPC and 80% of the negatively correlated metabolite pairs identified PPC have a higher absolute partial correlation, than the respective Pearson correlation. However, the magnitude of the changes is in the range of 0.01 to 0.08, similar to the observations from the *E. coli* data set.

Like in the analysis of the *E. coli* data set, we focused on a subset of fully annotated metabolite pairs. Of the 295 metabolite pairs of the TPC approach 150 were unambiguous annotated (Supplemental Table 8.3.10), whereas 773 out of the 1534 metabolite pairs of the PPC approach were unambiguous annotated (Supplemental Table 8.3.11). The difference in numbers of the TPC and PPC approach indicate a general tendency in the regulation towards post-transciptional regulation. This was already noted in the results of the original study in which the authors observed only a minor interconnection of the measured metabolites and transcripts. Further, they assume that this would change with a higher proportion of secondary metabolites, as primary metabolites have to react faster during external changes and are therefore mostly under post-transcriptional regulation [Caldana et al., 2011]. We next focus on specific examples of both approaches to show their capability to distinguish between both regulatory mechanism.

We start the investigation with the unique metabolite pairs identified with the TPC approach. We observed that the highest positive correlations are between amino acids and glycerol. In studies related to heat stress and heat tolerance it was shown that glycerol increased as a response to heat. Additionally, the studies showed an increase of amino acids as alanine, beta-alanine, leucine, isoleucine and aspartate [Kaplan et al., 2004]. We could report the pairs glycerol and isoleucine (Pearson correlation of 0.60 and partial correlation of 0.60), glycerol and leucine (Pearson correlation of 0.59 and partial correlation of 0.60) and glycerine and beta-alanine (Pearson correlation of 0.32 and partial correlation of 0.33). The measurements were done under different light and temperature conditions, including highlight and high temperatures. It is therefore realistic to assume, that we observe mild heat stress reactions. The regulation of heat stress response is reported to be completely under transcriptional regulation [Ohama et al., 2016] , which agrees with our findings.

Within the results of the PPC approach, we found amino acids correlating with each other. This observation is in agreement with previously published results, showing that the synthesizing pathways of most amino acids are under post-transcriptional regulation, more precisely under allosteric product inhibition [Less and Galili, 2008]. A well studied example is the branched-chain amino acid metabolism (BCAA) in which leucine, valine and isoleucine are synthesized. Each of these amino acids is reported several times within our PPC approach and forms pairs with other amino acids. Leucine and isoleucine are positively correlated to ornithine which is of interest as ornithine is a precursor of glutamte. Glutamate is involved in the synthesis of the BCAA amino acids. The reactions involved in these amino acid synthesis pathways are reported to be allosterically regulated [Binder, 2010].

Additionally, we found a relationship between skikimate and and related amino acids, as well as shikimate and sugars. Shikimate is a precursor to the amino acids tyrosine, phenylalanine and tryptophan. Shikimate is negatively correlated to pheylalanine (Pearson correlation of -0.42 and partial correlation of -0.39) and tyrosine (Pearson correlation of -0.59 and partial correlation of -0.57). Tryptophan was not reported within the uniquely identified metabolite pairs. At the same time shikimate is positively correlated to pyruvic acid (Pearson correlation of 0.67 and partial correlation of 0.64), fructose (Pearson correlation of 0.74 and partial correlation of 0.76), glucose (Pearson correlation of 0.78 and partial correlation of 0.80) and sucrose (Pearson correlation of 0.74 and partial correlation of 0.73). We therefore observed that metabolites upstream of shikimate (sugars) were positiveley correlated,

while downstream metabolites were negatively correlated. The pathway is partly feedback regulated meaning that the end products (amino acids) inhibit their production which explains the negative correlation. The sugars were positively correlated to shikimate as they are potential precursors [Tzin and Galili, 2010a,b].

In comparison to *E. coli*, we found more pairs with both approaches. The correlations of the TPC approach was higher than in *E. coli*. A similar situation was observed for the PPC approach. We observed more sugars and sugar derivatives in *E. coli*, whereas amino acids were mostly found with high positive correlation.

6.4 Discussion

In this study we proposed two approaches for a combined investigation of metabolic and transcriptomic data. The two proposed approaches are based on the concept of removing transcriptional information from metabolomic data, allowing us to categorize pairs of metabolites into transcriptionally or post-transcriptionally regulated. The developed approach *Transcriptional dependent Partial Correlation* allows the identification of transcriptionally regulated metabolites through a modified partial correlation approach, using PCs of the transcriptomic data as controlling variables. The second approach, *Post-transcriptional dependent Partial Correlation*, is based on a similar concept and it allows the identification of post-transcriptional regulation between pairs of metabolites.

The commonality of the investigated data sets is their focus on the change of central metabolites after perturbation or changing environmental conditions. It has been shown that in microorganisms the majority of primary metabolites are mainly regulated on the enzymatic level through feedback inhibition [Sanchez and Demain, 2008]. Further, the post-transcriptional regulation allows the organism to react faster to changes in the environment [Caldana et al., 2011]. The combination of these two criteria explain the larger amount of metabolite pairs found with the PPC approach, in comparison of the TPC approach.

The low coverage of correctly annotated metabolites in the data sets restricted our analysis to a smaller subset of metabolites. Nevertheless, the annotated metabolites were sufficient to obtain an overview over the potential of the approaches. We demonstrated that there is experimental evidence in the literature that the proposed approaches are capable of detecting differences in the association of metabolites, namely if the association is due to transcriptional or post-transcriptional effects. Moreover, we could show that our results agree with the findings from the study of Oliveira et al. [2015]. Metabolites that were reported to be post-transcriptionally regulated were also identified to participate in relationships identified by our PPC approach. We observed a similar situation with the transcriptionally associated metabolites, although we had to pool the reported metabolites from the "upstream" and "parallel" motif, as the TPC approach takes all transcriptional regulation mechanism into account.

While we observed a differentiation into pairs found by TPC and PPC, the detected partial correlation in each approach did not differ strongly from the found Pearson correlation (see Figure 6.3). The Pearson correlation captures most of the association already. Therefore, our approach does not strongly affects the correlation, but is a tool for categorizing the associations between metabolites. This claim is supported by two findings, the lack of overlap of metabolite pairs found with the two approaches in all three data sets and the low difference of the Pearson correlation and partial correlation for the identified metabolite pairs.

In the recent work of Bradley et al. [2008], they reported that the correlation between metabolites and transcripts depends on the experimental condition. The authors report that nearly no correlation was found when the correlation was investigated over multiple conditions, whereas high (positive or negative) correlations were observed if the conditions were investigated separately. Our approach aims to identify the general underlying relations between metabolites and if these originate from transcriptional regulation or post-transcriptional regulation. While the magnitude of the correlation is often of interest for many studies, our approach allows to gain further knowledge through the classification of the identified metabolite associations. Employing the approaches over multiple conditions allows us to give a general statement about the regulation associating pairs of metabolites.

A potential application for our proposed approaches is metabolic engineering. Metabolic engineering aims at enhancing certain important pathways leading to an overproduction of a metabolite of interest [Bailey, 1991; Nevoigt, 2008]. A frequently employed technique is the over-expression of genes associated with the metabolic pathway of interest. This technique has the disadvantage that the resulting phenotype (metabolite production) is difficult to predict and needs a strict monitoring for the validation. The results of the over-expression approaches might fall behind the expected yields of the metabolites. This shortcoming may be due to post-transcriptional regulation within the engineered pathway. Our method allows to investigate metabolic pathways before establishing over-expression lines and selecting metabolites and corresponding pathways which are mostly under transcriptional regulation, rather than post-transcriptional. This would allow biologists to focus their experiments to a smaller set of over-expression lines which would save both time and experimental resources.

Overall, we present here two approaches named TPC and PPC for investigating the prevalent regulatory mechanism of metabolite pairs. To our knowledge it is the first time that partial correlation is used to remove all transcriptional information from a metabolomic data set, removing not just the effect of a set of genes, but the majority of transcriptional regulation. This novel investigation methods will help to elucidate the complex regulatory mechanisms of metabolites while employing well known and established statistical methods.

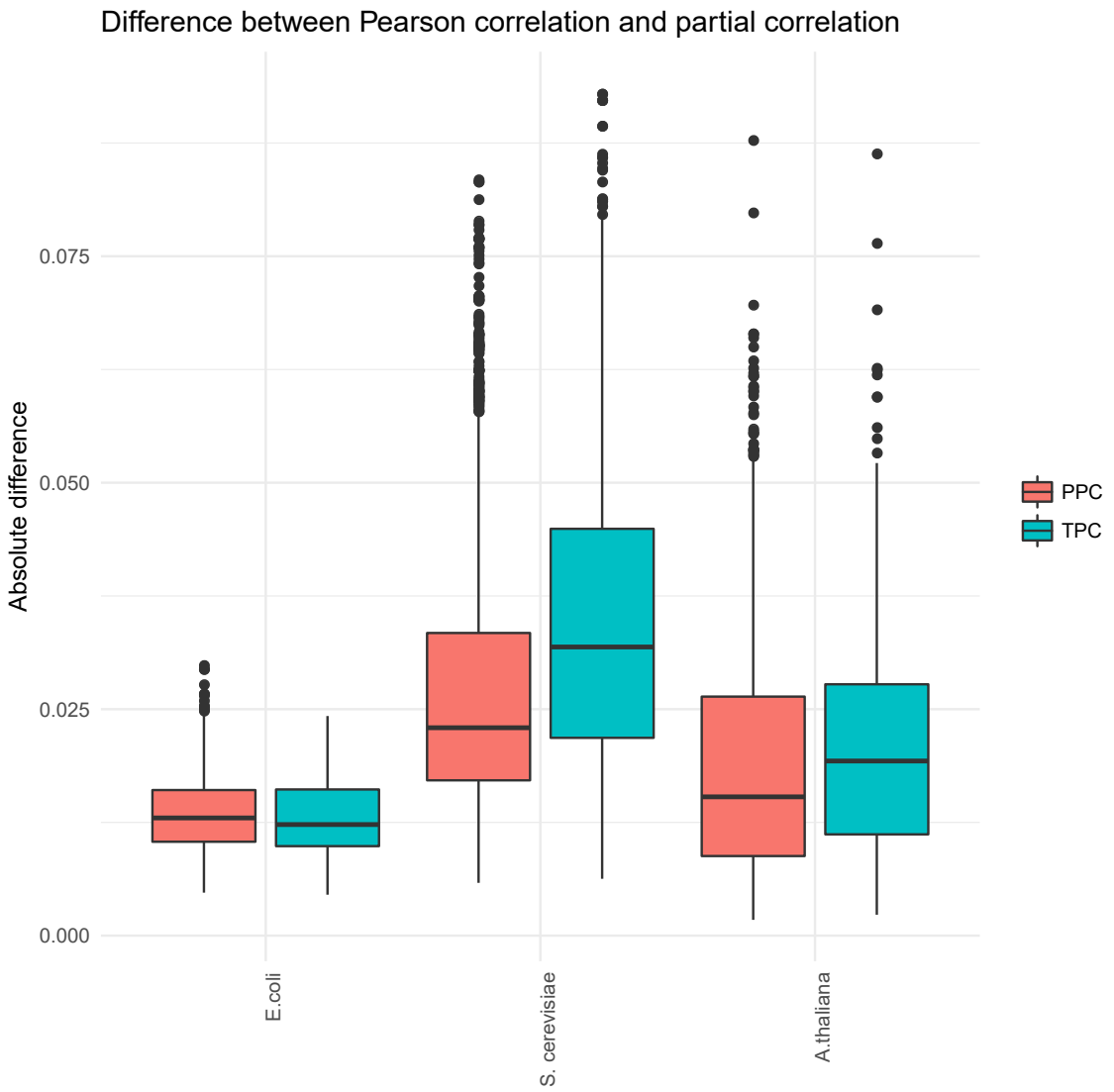

Figure 6.3: **Distribution of the absolute difference of Pearson and partial correlation.**
The boxplots show the absolute difference of the Pearson and partial correlations for each of the three organism (*E. coli*, *S. cerevisiae*, and *A. thaliana*) and the two approaches, respectively.

7 Discussion

The field of systems biology is facing an increase of available data as high-throughput technologies to measure transcriptomics, proteomics and metabolomics become available to most laboratories due to lower costs. While this increase of data brings advantages, it also increases the demand for new algorithmic and statistical approaches to analyze and integrate this data for the purpose of synthesizing new hypotheses and knowledge.

The existing high-throughput technologies differ with respect to the coverage of the monitored components. For instance, modern transcriptomic approaches yield a nearly complete overview of the transcripts present in a cell at a given time point and under specific experimental conditions. The technical improvements in the transcriptomics field have significantly expanded the detection capacity, increased the number of samples that can be measured in parallel, and have contributed to more precise measurements of lowly expressed genes [Heather and Chain, 2016]. In contrast, the available metabolomics technologies are not capable of detecting all metabolites present in a sample. This is due to the fact that the abundance of some metabolites can be below the detection limit of the currently available systems. Secondly, a majority of measured metabolites have not been identified [Fernie and Tohge, 2017]. The identification of metabolites from MS-spectra often requires the comparison of the m/z values and retention time with database information, which are still incomplete in regard to the large number of existing metabolites [Lai et al., 2018]. However, the combination of better extraction methods [Salem et al., 2017] with larger databases and tools to annotate unknown metabolites [Lai et al., 2018] will give the opportunity to overcome the current limitations.

In this closing chapter, I will briefly summarize the main findings of the three chapters of results. First, I investigated whether a previously established method from the observability theory can be validated investigating metabolite profiles from different model organisms. Secondly, I proposed a method allowing us to investigate coupling between different sets of metabolites, based on relative metabolite levels without the need of network information. The proposed approach facilitates the comparison of the extent of regulation present in the system. Finally, I proposed a method for the combined investigation of metabolomic and transcriptomic data which allowed a characterization of metabolite pairs whose association may be due to transcriptional or post-transcriptional regulation. As detailed discussions are provided within the respective studies, the following sections will put the findings into the broader context of systems biology.

7.1 Central metabolites as sufficient study objectives

In chapter 4, I investigated the applicability of a method from observability theory to metabolic networks. The approach was introduced in a previous study of Liu et al. [2013]. It allows us to identify so-called sensors in cellular (and metabolic) networks. Sensors contain the relevant information to reconstruct the internal state of the system of interest. Therefore, it is not necessary to have information of all components of the network since the sensors suffice to achieve this task. However, it was not shown what insight the approach provides in the context of metabolomics-centered studies.

Sensor metabolites can be determined by building the inference graph obtained from a given metabolomic network under the assumption that the reaction rates are described by mass action ki-

netics. A metabolite u (i.e. node) is connected by a directed edge to an other node v if the metabolite u occurs in the differential equation of v. The inference graph can then be decomposed into its strongly connected components (SCCs). A SCC is the maximal subgraph for which there are directed paths from every node to all others. If a SCC does not have an incoming edge, it is referred to as a root SCC. All metabolites within a root SCC are potential sensors. If a root SCC consists of more than one metabolite any of the metabolites within the SCC can serve as a sensor for that SCC. All remaining metabolites are non-sensor metabolites. The complete details of the approach are summarized in Figure 4.1.

However until the investigation performed in chapter 4, it has not been shown if and how these theoretical investigations can be used in metabolomics-centered studies. I could show that sensor metabolites exhibit higher correlations to each other than to the remaining non-sensor metabolites. In contrast, non-sensor metabolites are only weakly correlated to each other. This was reproducible for sensors and non-sensors predicted from two different Arabidopsis large-scale metabolic models. In order to calculate the correlation between sensors and non-sensors, I used publicly available metabolite profiles from *A. thaliana* grown under different light and temperature conditions [Caldana et al., 2011]. These results were supported by findings from a kinetic model of medium size used for simulating a synthetic data set in which the same relationship between sensor metabolites and non-sensor metabolites was observed.

My analyses revealed that most of the sensor metabolites from different SCCs were connected to the same non-root SCC. Therefore, the high correlation should be due to the position of the sensor metabolites in the network and their connection to the same non-root SCC. In order to test the hypothesis that the correlation is due to the position of the metabolites in the network, metabolites from the investigated data set were randomly assigned to be either a sensor or a non-sensor. The randomized sensors exhibit lower correlation values, revealing the network structure as a cause of correlation of the originally identified sensors.

The identification of sensor metabolites for the system of interest have the potential to establish more cost efficient and faster experimental setups. Instead of measuring and identifying the complete set of metabolites that can be detected with the MS-approach of choice, only the relevant sensor metabolites have to be analyzed. Measuring a smaller set of metabolites can speed up the investigation of the resulting data. Finally, if these metabolites are sufficient to reconstruct the information from the complete metabolic system, they can be used for a phenotypical description on the level of metabolites.

Especially in plant science, the effects of biotic and abiotic stresses are investigated and compared to a control environment as plants are sessile organism and can not evade stress conditions [Shulaev et al., 2008]. The study in chapter 4 could be extended to investigate the response to various stress conditions. First, the investigated approach can be used to perform comparisons between several different conditions on the same set of sensor metabolites. Sensors should show condition-specific behavior as they are sufficient to reconstruct the internal state of the system. A broader comparison over multiple stress conditions provides the opportunity to detect differences and similarities between stress conditions. In addition, the increase of the availability of high-throughput data allows the construction of condition-specific models [Estévez and Nikoloski, 2014; Vlassis et al., 2014]. The approaches rely on transcriptomic and metabolomic measurements constraining the network to include metabolic reactions which are active under the investigated condition [Estévez and Nikoloski, 2014]. The resulting condition-specific networks are shown to be in agreement with physiological data [Becker and Palsson, 2008]. These condition-specific networks can be used to estimate condition-specific sensors which would allow to reveal additional stress induced differences on the metabolic level.

However, it should be noted that the approach is highly dependent on the quality of the available metabolomic networks from which the information about the sensor metabolites are gained, as only complete and high quality networks ensure the discovery of all relevant sensor metabolites. While advances have been made in generating and optimizing genome-scale metabolic networks for a broad set of organisms [Henry et al., 2010; Kim et al., 2012], the networks might contain erroneous reactions [Fritzemeier et al., 2017]. Therefore, identified sensor metabolites should be carefully investigated with respect to gathered data profiles. This can be done through verifying that the correlations within sensors is higher than within non-sensor metabolites, as shown in chapter 4. In addition, the discussed approach from observability theory is based on mass action kinetics. Mass action kinetics assumes an elementary reaction whose reactant molecules (metabolites) collide. Further, the system has to be well-mixed so that the probability that they meet is proportional to the product of their concentrations. These assumptions are not necessarily valid in biological systems. In addition, mass action cannot model regulatory mechanism occurring at the level of enzymes, such as inhibition or cooperativity. Therefore, extensions of mass action could be used to integrate regulation into metabolic networks which would subsequently allow a more precise identification of sensor metabolites.

7.2 Reaction coupling - network information in metabolomic data

In chapter 5, I proposed a novel approach to estimate the number of coupled metabolite sets to reflect the coupling of reaction rates. The novelty of the approach is that it is purely data-driven. The approach is based on the concept that systems sense and respond to environmental perturbations while achieving normal functionality and assuming that elementary biochemical reactions can be modeled via mass action kinetics. This can be expressed with the equation:

$$\frac{k_p \prod_i x_i^{\alpha_{ip}}}{k_q \prod_i x_i^{\alpha_{iq}}} = \frac{k_p}{k_q} \prod_i x_i^{\alpha_{ip}-\alpha_{iq}} = \gamma_{pq} \tag{7.1}$$

Two reactions p and q have coupled rates if their ratio is a constant γ_{pq}. The ratio solely depends on the metabolite levels, represented by x_i, as the stoichiometric coefficients α_{ip} and α_{iq} and the enzymatic parameters k_p and k_q are assumed to be constant. Therefore, the combination of the metabolite levels related to these two coupled reactions should be highly correlated, as they have to change in concordance. If the assumptions holds, one can estimate the degree of coupled reaction rates through the calculation of stoichiometric correlations, termed stoichiometric correlation analysis (SCA). SCA is calculated for pairs, triples and quadruples of metabolites. The reasoning behind the three cases is that most metabolic reactions have either one or two substrates. The calculation of pairs is performed using Pearson correlation $r(log(x_i), log(x_j))$, for all couples $1 \leq i \neq j \leq n$ of metabolic profiles x. The calculation of triples is implemented through finding $a, b \in \{1,2,3,4\}$ maximizing the Pearson correlation between $a\, log(x_i) + b\, log(x_j)$ and $log(x_k)$. Similarly, the calculation of quadruples is performed through finding $a, b, c, d \in \{1,2,3,4\}$ maximizing the Pearson correlation between $a\, log(x_i) + b\, log(x_j)$ and $c\, log(x_k) + d\, log(x_l)$. The most common stoichiometric coefficients in metabolite reactions can be represented by $a, b, c, d \in \{1,2,3,4\}$. Employing this approach, I was able to show that the stringent response in *A. thaliana* is more controlled on the level of reactions in comparison to *E. coli* based on the same set of metabolites from the TCA cycle and amino acids. In addition, I could show that

through the process of domestication the number of stoichiometric correlations decreases indicating that there is a loss of regulatory control in the domesticated species. The loss of regulatory control was independently reproduced by employing data from domesticated and wild wheat and tomato species.

The study presented in chapter 5 is the first to investigate if the coupling of reaction rates is reflected in metabolic profiles. The investigation of reaction coupling has been performed so far on metabolic networks using flux coupling analysis (FCA) [Burgard et al., 2004]. In FCA, the minimum and maximum ratios of all combination of reactions in a network are calculated. Based on the relationship between the resulting values uncoupled, directionally, partially and fully coupled reactions are determined [Burgard et al., 2004]. In the context of my study, the comparison to fully coupled reactions is most relevant. A pair of fully coupled reactions is found if the maximum and minimum ratio of the two reactions is a constant [Burgard et al., 2004], similar to the description of coupled reaction rates in equation 7.1. Within the presented study in chapter 5, I compared the total number of observed stoichiometric correlations between two species. While this strategy is advantageous for investigating changes during domestication or differences between species, it did not account for an in-depth analysis of the identified metabolites within the pairs, triples and quadruples. Therefore, the substrate metabolites of two fully coupled reactions assigned with FCA could be further investigated with SCA. However, FCA requires to have access to the underlying metabolic network. The combination of FCA and SCA would allow us to validate the results obtained with FCA. In addition, the combination of both would give the opportunity to additionally validate genome-scale metabolic networks. The reasoning behind this is that FCA is sensitive to missing reactions in the network [Marashi and Bockmayr, 2011]. A high stoichiometric correlation for a pair, triple or quadruple of metabolites that cannot be detected with FCA might indicate an erroneous part of the network used or the reaction kinetic deviates from mass action.
Nevertheless, the above described limitations of the mass action kinetics holds similarly for the estimations of stoichiometric correlations. Therefore, the usage of an extension of the mass action kinetics provides the possibility to identify coupled reaction rates while accounting for enzymatic regulations.

7.3 Integration of data types - towards the investigation of regulatory effects

The regulatory mechanisms between transcription and metabolism have been intensively studied before employing a variety of different approaches. These approaches include Pearson and Spearman correlation to elucidate single gene to metabolite association [Gibon et al., 2006; Hannah et al., 2010], regression-based approaches for the investigation of multiple genes and one metabolite [Auslander et al., 2016] and multiple genes and multiple metabolites with approaches such as Partial Least Squares (PLS) [Bylesjö et al., 2007] and canonical correlation analysis (CCA) [Jozefczuk et al., 2010]. All mentioned studies focus on the transcriptional control on metabolism while not including regulatory effects from the post-transcriptional level.

In chapter 6, I presented two novel approaches for the combined investigation of transcriptomic and metabolomic data, termed *Transcriptional dependent Partial Correlation* (TPC) and *Post-transcriptional dependent Partial Correlation* (PPC). These two methods allow for the categorization of metabolite pairs as either being associated due to transcriptional or post-transcriptional regulation. Each of the two approaches is composed of four steps. The first two steps are the calculation of the

first p principal components (PCs) of the transcriptomic data and the determination of all metabolite pairs with a significant Pearson correlation. The third step differs between TPC and PPC. For the TPC approach, all metabolite pairs with a non-significant partial correlation upon removal of the p PCs (the controlling variables) are determined. In contrast, in the PPC approach all significant partial correlations are determined upon removal of the controlling variables. In the fourth step, the pairs of metabolites that show a significant difference between their Pearson and partial correlation values obtained in step two and three are retained. The reason for this construction of the approach is the following: if the removal of the significant PCs leads to a non-significant partial correlation between two metabolites, their association may be due to transcriptional regulation. In contrast, if the correlation remains upon removal of the transcriptional effect, the observed association may be due to post-transcriptional regulation of the two metabolites. In addition, the significant difference to the observed Pearson correlation is employed to ensure that the observed partial correlation is due to transcriptional effects in the case of the TPC approach and could not be found with Pearson correlation alone. The same is done for the PPC approach to ensure that the correlation is due to post-transcriptional effects.

In the study, I was able to confirm the characterization of pairs associated due to transcriptional or post-transcriptional regulation by comparing the obtained results with previous published experimental results. In general, I found a larger amount of metabolite pairs with the PPC approach in comparison of the TPC approach. The reasoning behind this observation is that the analyzed data came from studies in which the change of central metabolites after perturbation or changing environmental conditions were investigated. It has been shown that feedback inhibition of enzymes and post-transcriptional regulation allows organisms to react faster to environmental changes [Sanchez and Demain, 2008; Caldana et al., 2011]. Therefore, the high number of metabolites associated due to post-transcriptional regulation found with the PPC approach are in agreement with previous studies. Moreover, I was able to identify pairs of amino acids with the PPC approach in the *A. thaliana* and *E. coli* data sets which are known to be mainly regulated through allosteric product inhibition [Less and Galili, 2008; Sanchez and Demain, 2008].

In addition, I could show that the classification coincide with results from the study of Oliveira et al. [2015]. They used a Bayesian inference framework approach to investigate the effects of TORC1 during nitrogen supply shift (upshift and downshift) as well as the treatment with Rapamycin in *S. cerevisiae*. While their approach allows for the characterization of metabolite pairs in relation to TORC1, the TPC and PPC approaches are capable of identifying additional pairs of metabolites whose association is due to regulatory mechanism that are independent of TORC1.

The proposed approaches in chapter 6 extend the investigation of regulation of metabolites twofold. First, it allow for the investigation of the complete set of measurable transcripts and all measurable metabolites. Therefore, it can give an even more comprehensive overview over the transcriptional influence on metabolism. Secondly, the approach allows us to study post-transcriptional regulation between metabolite pairs to further understand the regulation of metabolism in its entirety. The multitude of post-translational modifications of enzymes performed by kinases and phosphatases [Gonçalves et al., 2017], as well as feedback regulation, have major impacts on metabolism [Friso and van Wijk, 2015]. This large amount of possible regulatory effects makes it challenging to investigate all of them in detail. The PPC approach allows to estimate post-transcriptional regulation between metabolites from a purely data-driven perspective. While the PPC approach only detects post-transcriptional regulation as a cause, it can not estimate the exact regulatory mechanism. However, the approach could be used to pre-select sets of metabolites for additional investigations to specifically

determine the regulatory mechanisms.

7.4 System-wide investigation of regulatory mechanism on metabolism

While metabolite levels are affected through the sum of regulatory mechanisms influencing transcription and translation which in turn influences enzyme levels, they themselves have regulatory properties and affect transcription and gene expression via DNA and histone modifications [Donati et al., 2018; Yugi and Kuroda, 2017]. Further, metabolites alter post-translational modifications of enzymes and accordingly their activity [Donati et al., 2018; Yugi and Kuroda, 2017]. Therefore, the multitude of possible regulatory effects renders any experimental analysis on a genome-scale level cumbersome [Gerosa et al., 2015]. In order to investigate system-wide regulatory mechanisms that shape metabolism or are affected by metabolites, several computational methods have been employed to date. Beside the usage of mass action kinetics to model the behavior of biological systems, extensions such as a power-law formulation have also been used. While mass action was initially used to describe chemical reaction rates, it has been extended to describe metabolic systems [Voit et al., 2015]. In general, mass action kinetics can be used to describe reaction rates of elementary reaction in well-mixed systems if the number of molecules is high and the amount enzymes is not rate limiting [Sayikli and Bagci, 2011]. In addition, the usage of mass action kinetics is motivated by the fact that it requires less parameters to describe a metabolic reaction than for example Michaelis-Menten kinetics [Du et al., 2016]. However, the formulation of the mass action kinetics does not allow to integrate regulatory effects. To overcome this limitation while at the same time not increasing the complexity drastically, extensions of mass action have been used.

A power-law formulation based on mass action was proposed by Savageau [1988] and can be used to describe the effect of multiple inputs contributing to an output. The equation 7.2 specifies the generic power-law representation, with α_i being the non-negative rate constant, g_{ij} decides if the term is activating (positive) or inhibiting (negative), I_j being the input into the system and the O_i output [Savageau, 1988; Voit et al., 2015].

$$O_i = \alpha_i \prod_{j=1}^{n} I_j^{g_{ij}} \tag{7.2}$$

The transcriptional response of an organism can be modeled with such a power-law formulation, accounting for the general expression machinery and regulatory affects of transcription factors, describing the promoter activity [Kochanowski et al., 2017]. The regulatory impact of metabolites on the specific transcriptional regulation can than be investigated, upon subtracting the global regulation. In the case of the investigation of the *E. coli* central carbon metabolism, the approach showed that the majority of the regulation of gene expression can be attributed to global regulatory mechanisms while single metabolites modulate the specific regulation [Kochanowski et al., 2017]. In addition, Gerosa et al. [2015] investigated regulatory mechanisms during steady-states transitions based on a power-law formulation. The assumption is that metabolites and transcripts regulating the transitions should display significant differences between these states. The approach requires access to ^{13}C-flux measurements, metabolite levels and transcript measurements to estimate enzyme and transcription factor

abundance. The power-law formulation is used to describe the transcriptional, thermodynamic and substrate regulation on the metabolic flux change between two conditions. The thereby estimated regulation coefficients indicate that during carbon source changes transcriptional and substrate regulation contributed equally to the changing fluxes. However, the method does not account for inhibition, post-translational modification or allosteric regulation.

Beside the transcriptional regulation, metabolites shape the metabolic fluxes through the regulation of enzymes, either as the substrates or as allosteric regulators of the enzymes [Reznik et al., 2017]. Therefore, having detailed knowledge which metabolites affect (activate or inhibit) an enzyme is beneficial for any system-wide investigation. Regulatory metabolites can be investigated through a network-based approach employing genome-scale metabolic networks and enzyme database information [Reznik et al., 2017]. The combination of these two data sources allows to invest metabolite feedback regulation and the distinction between the roles of a metabolite, as either substrate or inhibitor, for any organism with a genome-scale model and sufficient information within the BRENDA database. While the approach allows to perform investigations on a genome-scale level, it is purely based on previous reported information and cannot be used to invest unknown regulatory mechanisms. Moreover, the enzymatic parameters have been mainly investigated in *in vitro* experiments which do not necessarily resemble the *in vivo* conditions [Schwender and Junker, 2009].

This shortcoming is acknowledged by the approach from Hackett et al. [2016] estimating if a reaction follows Michaelis-Menten kinetics. The approach fits experimental data (enzyme and metabolite concentration) following Michaelis-Menten kinetics to metabolic fluxes. If a reaction is found to deviate from Michaelis-Menten kinetics, regulatory metabolites are included to enhance the fit to the fluxes. Therefore, the approach is capable of detecting previously unknown allosteric regulation and the involved metabolites on a genome-scale level, provided sufficient experimental data are available [Hackett et al., 2016]. However, the quality of the estimations could be increased by using ^{13}C-flux measurements instead of fluxes estimated from constraint-based modeling. The quality of the used stoichiometric network will subsequently influence the estimation of the enzyme kinetics. Nevertheless, the approach revealed that metabolite concentrations are determinant of metabolic reaction rates while enzyme concentrations have a minor contribution. This is in agreement with other studies which found that enzymes of the central metabolism are present in a high abundance and the fine-tuning of the fluxes is performed by allosteric regulation [Donati et al., 2018; O'Brien et al., 2016]. Similar results have been found for the metabolites and the enzymes of the Calvin-Benson cycle (CBC). Several enzymes of the CBC, including including GAPDH, aldolase, and TPI are present in high concentrations [Sulpice et al., 2010] and are above the concentrations of their respective substrate metabolites [Mettler et al., 2014]. The relative low metabolite concentrations result in free capacities of the enzymes. This allows to increase the flux rapidly upon the availability of substrate metabolites induced through high light intensities [Mettler et al., 2014]. However, the substrate concentrations of the enzymes Rubisco, FBPase and SBPase are close to the K_M value of the enzymes indicating that an increase of the reaction rate has to be performed through changes on the transcriptional or post-transcriptional level [Mettler et al., 2014]. This further points out the complexity of the interplay of the different level of metabolic regulation.

The studies above already provide insights into regulatory mechanisms of the metabolome, partly employing extensions of the mass action kinetic. To additionally elucidate the metabolic regulation, additional investigations based on the power-law description could be used. There exists two representation, the S-system formulation (equation 7.3) and generalized mass action (GMA) kinetics (equation

7.4), with x_i being the metabolite concentration, α_i and β_i being the rate constant of total influxes and outfluxes and $g_{ij(k)}$ and $h_{ij(k)}$ being kinetic orders of influxes and outfluxes [Sriyudthsak et al., 2016].

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{n} x_j^{g_{ij}} - \beta_i \prod_{j=1}^{m} x_j^{h_{ij}} \tag{7.3}$$

$$\frac{dx_i}{dt} = \sum_{j=1}^{p} \alpha_{ij} \prod_{k=1}^{n} x_k^{g_{ijk}} - \sum_{j=1}^{q} \beta_{ij} \prod_{k=1}^{m} x_j^{h_{ijk}} \tag{7.4}$$

Within the S-system formulation, all influxes are collected and their sum is represented as a single power-law term. The same is done for all effluxes resulting in a single difference of the two power-law terms. In contrast, the GMA formulation represents each metabolic reaction as one power-law term. Therefore, the right hand-hand side of the differential equations consist of the difference between the sum of power-law terms for all influxes and the sum of power-law terms for all effluxes [Voit et al., 2015]. The difference to mass action description is the inclusion of the kinetic order of the reaction allowing to include the influence of regulating (enhancing or inhibiting) metabolites.

As both S-systems and GMA allows to formulate the change of metabolite levels with ordinary differential equations (ODE), the graphical approach from Liu et al. [2013] to generate the inference graphs can be used with this extension of the mass action kinetic. In addition to substrate and product metabolites, the power-law formulation can integrate regulatory metabolites into the inference graph. With these being part of the network, allosterically regulating metabolites can potentially be identified as sensors. Therefore, metabolic phenotyping can be extended towards these regulating metabolites which will potentially emphasize the role of allosteric regulation within metabolism. However, in order to construct the inference graphs, stoichiometric metabolic networks are not sufficient, as these do not contain information about regulating metabolites; instead, kinetic models could be used. As the graphical approach from Liu et al. [2013] does only consider the existence of a metabolite in the ODE, no additional knowledge about the rate constant and the kinetic order are needed. Nevertheless, the available kinetic models are smaller than stoichiometric networks [Stanford et al., 2013]. Therefore, using the smaller kinetic networks for the graphical approach would limit the identification of sensor metabolites to sub-parts of the metabolome. Increasing the size and quality of kinetic networks will subsequently allow a system-wide overview of allosteric regulatory mechanisms.

In addition, the SCA approach (chapter 5) could be extended towards S-systems kinetics. Similar, to the mass action description, the ratio of two reactions rates formulated with the power-law kinetic would be a constant if the two reactions are coupled. Therefore, the substrate metabolites of two reactions would exhibit high correlation values close to one. A modified SCA would allow to integrate regulatory metabolites into the investigation and therefore account for allosteric regulation within reaction coupling. However, this will automatically increase the computational effort. The reasoning behind is to add regulatory metabolites in addition to the combination of two, three and four metabolites (pairs, triples and quadruples). In general, more than one metabolite could participate in an allosteric regulation of one reaction. Even in the simplest case of only one additional regulatory metabolite, the number of possible combinations that are needed to be calculated would rise tremendously. Currently, the implementation of the SCA calculates possible stoichiometric values with which the metabolites participate in the reaction. These are limited to the set of stoichiometric values of $a, b, c, d \in \{1, 2, 3, 4\}$. In contrast, a S-system kinetic implementation would need to estimate the kinetic order which can be positive or negative non-integer numbers. This larger number of possible kinetic order values would

even further increase the running time. In order to reduce the calculation time, one should limit the SCA to certain pathways of interest. First of all, this would reduce the number of possible metabolite combinations. In addition, the search space for the kinetic order values could be limited for specific metabolites to be either positive (activating) or negative (inhibiting) through the integration of additional literature information. Even though a power-law based implementation of SCA would be more computational demanding than the mass action version, it would allow to investigate allosteric regulation from a purely data driven perspective, using already existing data sets.

In order to unravel the multitude of regulatory mechanisms influencing metabolite levels and metabolic reaction rates, a combination of approaches is most likely to be successful. The approaches described above can be used to identify transcriptional regulation, regulatory metabolites and enzymatic kinetic parameters. The approaches discussed in this thesis can further contribute to elucidating the complex regulatory machinery by integrating regulation through coupling of reaction rates and the classification of the prevailing regulatory mechanism between metabolite pairs.

8 Appendices

8.1 Appendix: Observability of plant metabolic networks is reflected in the correlation of metabolic profiles

8.1.1 Supplemental Figures

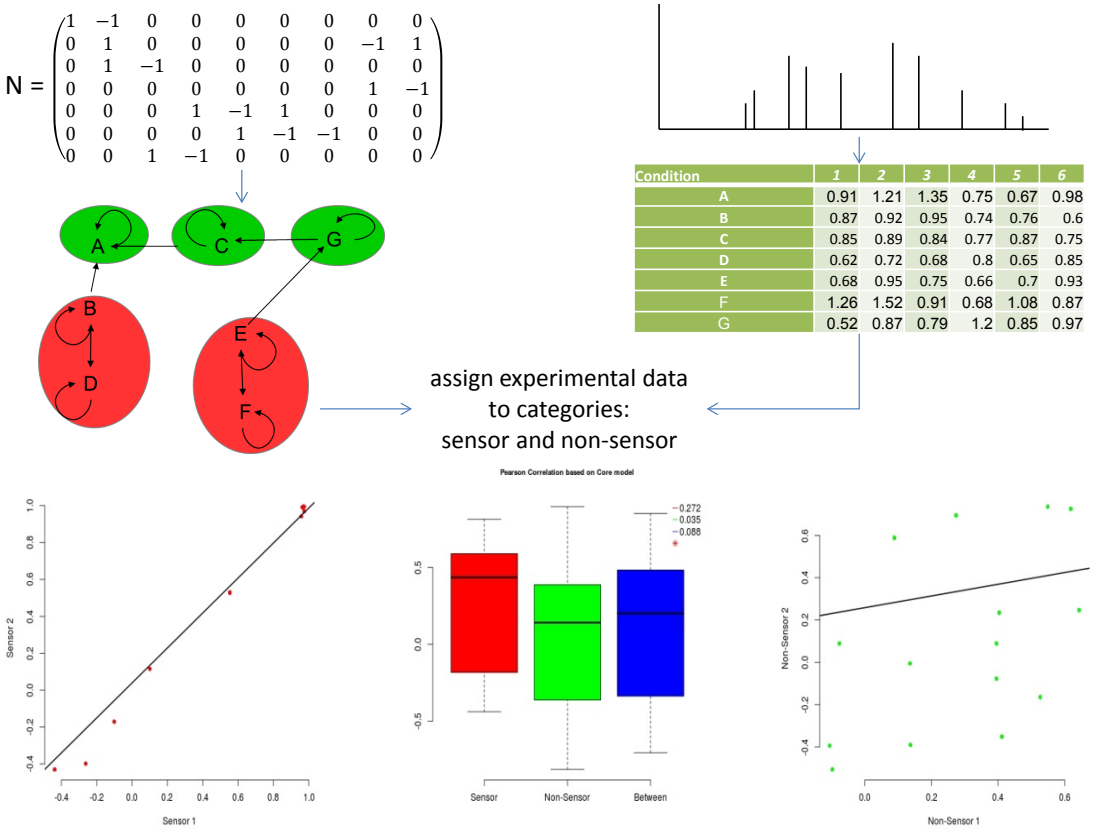

Figure 8.1.1: **Schematic overview of the procedure to compare sensor and non-sensor metabolites.** The stoichiometric matrix form the metabolic network is used as an input for the sensor identification algorithm (see Figure 4.1 for the approach). Experimental data from components A - G are combined with the identified sensors and non-sensors. We determine and compare standard deviations, coefficients of variation and Pearson correlation coefficients for sensor and non-sensor metabolites.

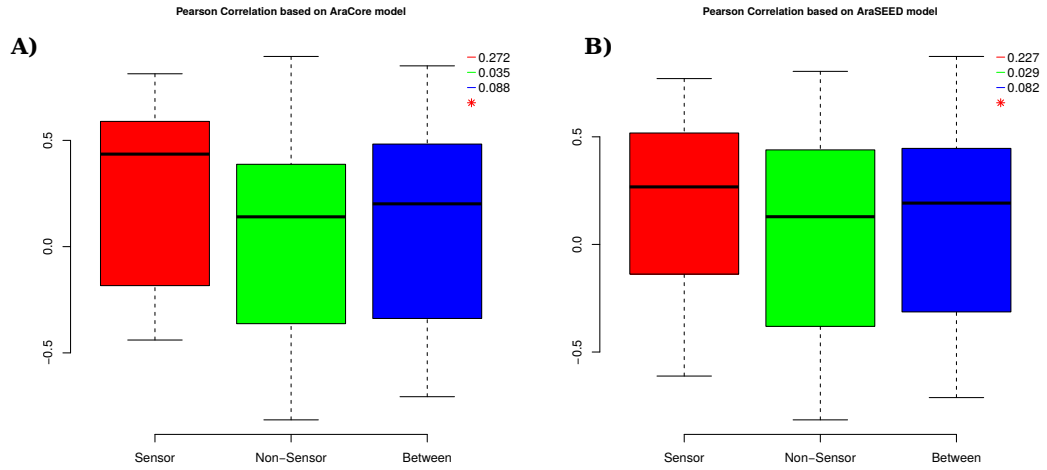

Figure 8.1.2: **Analysis of metabolite profiles of multiple Arabidopsis accessions.**
Comparison of Pearson Correlation values of sensors (red), non-sensors (green) and between sensors and non-sensors (blue). Data from Sulpice et al. [2013]. Metabolic data profiles from three different conditions of nitrogen supply and photoperiod were used. Number above the plot represent correspondence mean values. A red asterisk represent a significant difference in means of sensors and non-sensors ($\alpha = 0.05$). **A)** Using sensor and non-sensor information from the Arabidopsis core model. **B)** Using sensor and non-sensor information from the AraSEED model

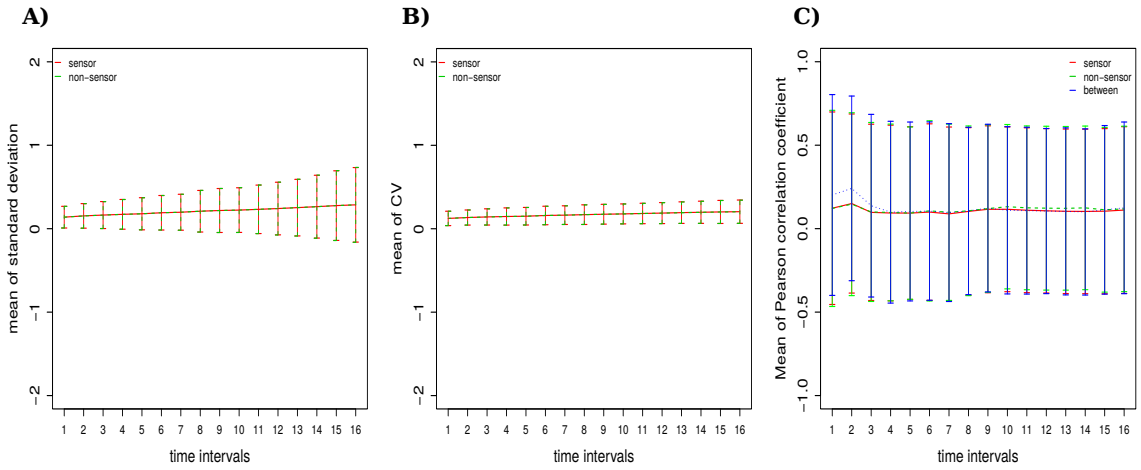

Figure 8.1.3: **Statistical comparison after randomizing the sensors and non-sensors.**
The x axis represents the investigated time interval, from 1 to 16. The y axis represents values for the three statistics, respectively: **A)** standard deviation, **B)** coefficient of variation, **C)** Pearson correlation of sensors and non-sensors. Red line corresponds to the values for the statistics between sensor metabolites, while green line corresponds to values between non-sensor metabolites. The blue line in (C) is used for the correlation between sensors and non-sensors. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and non-sensors. Bars represent the range of ± 1SD from the mean value, for 500 randomizations.

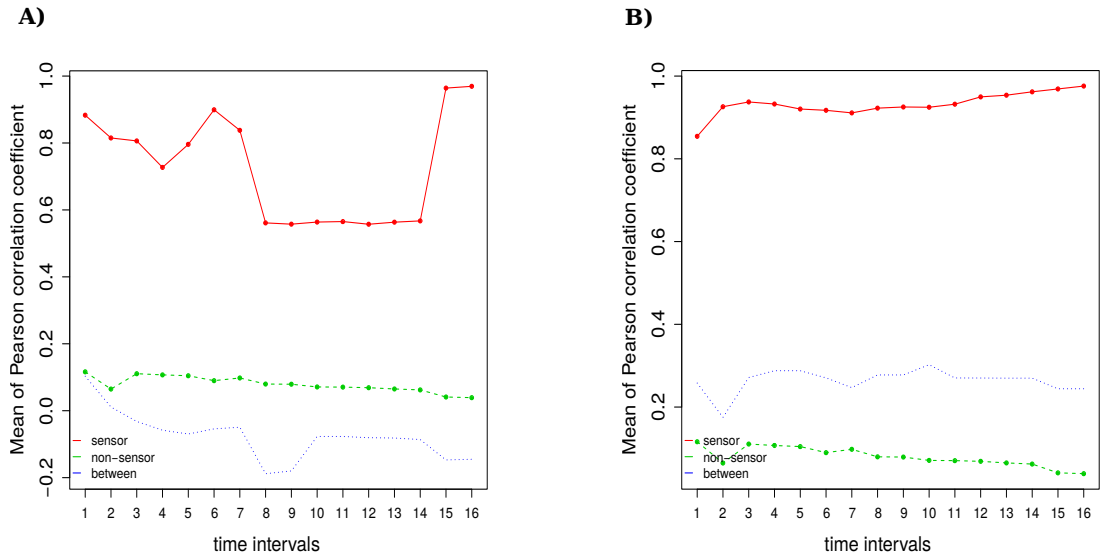

Figure 8.1.4: **Comparison of Pearson correlation of sensors and non-sensors of the AraSEED model connected or disconnected to the largest non-root SCC**
The x axis represents the investigated time interval, from 1 to 16. The y axis represents values for the Pearson correlation of sensors and non-sensors. Red line corresponds to the values for the statistics between sensor metabolites, while green line corresponds to values between non-sensor metabolites. The blue line in is used for the correlation between sensors and non-sensors. A dot on the line indicates a significant difference at level $\alpha = 0.05$ between sensor and non-sensors. **A)** Plot for sensors of the AraSEED model connected to the largest non-root SCC. **(B)** Plot for sensors of the AraSEED model not connected to the largest non-root SCC (fructose and glucose).

8.1.2 Additional files and tables

Additional supplemental files can be found at: https://owncloud.mpimp-golm.mpg.de/index.php/s/qVeeHs4Q7p73d5u

The file Paper1_Observability_Supplemental_Tables.xlsx contains the following tables:

Supplemental Table 8.1.1 List of identified sensor metabolites in the AraCORE model with their corresponding model ID and root sensor identifier

Supplemental Table 8.1.2 List of identified sensor metabolites in the AraCORE model without biomass reaction with their corresponding model ID and root sensor identifier

Supplemental Table 8.1.3 List of identified sensor metabolites in the AraSEED model with their corresponding model ID and root sensor identifier

Supplemental Table 8.1.4 List of identified sensor metabolites in the Kinetic model with their corresponding model ID and root sensor identifier

Supplemental Table 8.1.5 List of mapped sensor and non-sensor metabolites from the AraCORE model to the metabolite data of Caldana et al. [2011]

Supplemental Table 8.1.6 List of mapped sensor and non-sensor metabolites from the AraSEED model to the metabolite data of Caldana et al. [2011]

Supplemental Table 8.1.7 Overview of the used models with reaction and metabolite numbers

The file Manuscript_Observability_Supplemental_Material_Kinetic_Model.xlsx contains the kinetic model which was used for the performed day and night simulations as an Excel file. The file contains the stoichiometric matrix, the list of metabolites, the list of parameters and the list of reactions.

The file Manuscript_Observability_Supplemental_Material_Kinetic_Model_MATLAB.mat contains the kinetic model, which was used for the performed day and night simulations as a matlab file.

8.2 Appendix: Stoichiometric correlation analysis: principles of metabolic functionality from metabolomics data

8.2.1 Additional files and tables

Additional supplemental files can be found at: https://owncloud.mpimp-golm.mpg.de/index.php/s/qVeeHs4Q7p73d5u

Supplemental Table 8.2.1 Comparison between the number of significant stoichiometric correlation at four thresholds (0.8, 0.85, 0.9 and 0.95) for single step mass action and enzyme-metabolite

Supplemental Table 8.2.2 Quintiles of the stoichiometric correlations for all investigated species

Supplemental Table 8.2.3 Comparison between the number of significant stoichiometric correlation at three different thresholds (0.8, 0.85 and 0.9) for *E. coli* and *A. thaliana* and common significantly correlated metabolite pairs, triples and quadruples

Supplemental Table 8.2.4 Coupling degrees of metabolites for the comparison of *E. coli* and *A. thaliana*, *T. durum*, *T. dicoccoide*s and *T. dioccum,* M82 and wild tomato and *F. ananassa* and *F. vesca* at the thresholds 0.8, 0.85 and 0.9

Supplemental Table 8.2.5 Comparison between the number of significant stoichiometric correlation at the threshold of 0.80 for *T. durum*, *T. dicoccoide*s and *T. dioccum* common significantly correlated metabolite pairs, triples and quadruples

Supplemental Table 8.2.6 Comparison between the number of significant stoichiometric correlation at the threshold of 0.85 for *T. durum*, *T. dicoccoide*s and *T. dioccum* common significantly correlated metabolite pairs, triples and quadruples

Supplemental Table 8.2.7 Comparison between the number of significant stoichiometric correlation at the threshold of 0.90 for *T. durum*, *T. dicoccoide*s and *T. dioccum* common significantly correlated metabolite pairs, triples and quadruples

Supplemental Table 8.2.8 Comparison between the number of significant stoichiometric correlation at three different thresholds (0.8, 0.85 and 0.9) for M82 and wildtype tomato and common significantly correlated metabolite pairs, triples and quadruples

Supplemental Table 8.2.9 Comparison between the number of significant stoichiometric correlation at three different thresholds (0.8, 0.85 and 0.9) for *F. ananassa* and *F. vesca* and common significantly correlated metabolite pairs, triples and quadruples

The folder Manuscript_SCA_GitHub_code contains the code and one example of the publication and can also be found at: https://github.com/KSchwahn/Stoichiometric-correlation

8.3 Appendix: Data reduction approaches for dissecting transcriptional effects on metabolism

8.3.1 Supplemental Figures

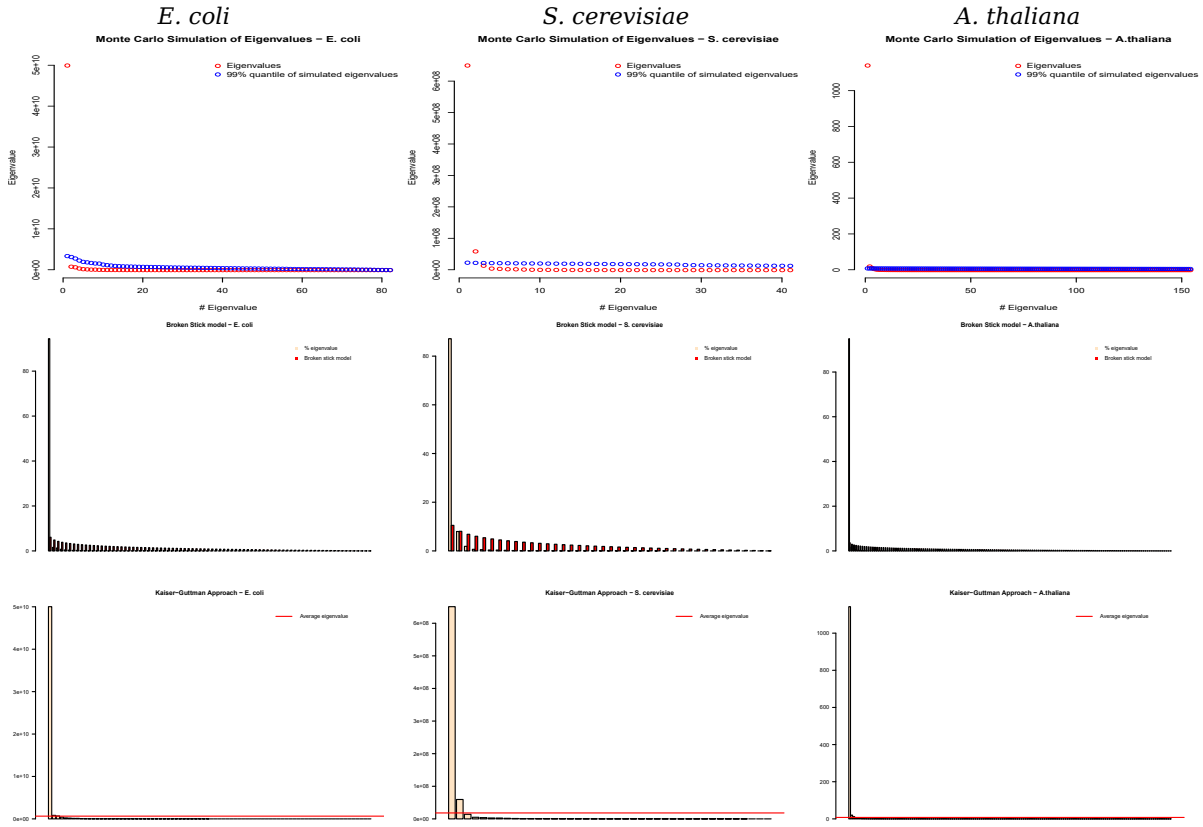

Figure 8.3.5: **Overview of significant PCs of transcriptomic data of the three species *E. coli, S. cerevisiae* and *A. thaliana*.**
Investigation of the number of significant PCs for the three species *E. coli, S. cerevisiae* and *A. thaliana* with Horn's parallel analysis, the Kaiser-Guttman approach and the Broken Stick model

8.3.2 Additional files and tables

Additional supplemental files can be found at: https://owncloud.mpimp-golm.mpg.de/index.php/s/qVeeHs4Q7p73d5u

Supplemental Table 8.3.1 Subset of the metabolomic and transcriptomic data from Oliveira et al. [2015] used in the study.

Supplemental Table 8.3.2 List of metabolite pairs identified with the TPC approach with the *E. coli* data set

Supplemental Table 8.3.3 List of metabolite pairs identified with the PPC approach with the *E. coli* data set

Supplemental Table 8.3.4 List of fully annotated metabolite pairs identified with the TPC approach with the *E. coli* data set

Supplemental Table 8.3.5 List of fully annotated metabolite pairs identified with the PPC approach with the *E. coli* data set

Supplemental Table 8.3.6 List of fully annotated metabolite pairs identified with the TPC approach with the *S. cerevisiae* data set

Supplemental Table 8.3.7 List of fully annotated metabolite pairs identified with the PPC approach with the *S. cerevisiae* data set

Supplemental Table 8.3.8 List of metabolite pairs identified with the TPC approach with the *A. thaliana* data set

Supplemental Table 8.3.9 List of metabolite pairs identified with the PPC approach with the *A. thaliana* data set

Supplemental Table 8.3.10 List of fully annotated metabolite pairs identified with the TPC approach with the *A. thaliana* data set

Supplemental Table 8.3.11 List of fully annotated metabolite pairs identified with the PPC approach with the *A. thaliana* data set

9 Bibliography

Adibi, M., Yoshida, S., Weijers, D., and Fleck, C. (2016). Centering the organizing center in the *Arabidopsis thaliana* shoot apical meristem by a combination of cytokinin signaling and self-organization. *PLOS ONE*, 11(2):1–28.

Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405.

Alonso, A., Marsal, S., and Julià, A. (2015). Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in Bioengineering and Biotechnology*, 3(23).

Antoniewicz, M. R. (2015). Methods and advances in metabolic flux analysis: a mini-review. *Journal of Industrial Microbiology & Biotechnology*, 42(3):317–325.

Araus, J. L. and Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science*, 19(1):52–61.

Arnold, A. and Nikoloski, Z. (2011). A quantitative comparison of Calvin-Benson cycle models. *Trends in Plant Sciene*, 16(12):676–83.

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

Auslander, N., Yizhak, K., Weinstock, A., Budhu, A., Tang, W., Wang, X. W., Ambs, S., and Ruppin, E. (2016). A joint analysis of transcriptomic and metabolomic data uncovers enhanced enzyme-metabolite coupling in breast cancer. *Scientific Reports*, 6:29662.

Awad, H., Khamis, M. M., and El-Aneed, A. (2015). Mass spectrometry, review of the basics: ionization. *Applied Spectroscopy Reviews*, 50(2):158–175.

Baba, K., Shibata, R., and Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zeawland Journal of Statistics*, 46(4):657–664.

Baginsky, S. (2009). Plant proteomics: concepts, applications, and novel strategies for data interpretation. *Mass Spectrometry Reviews*, 28(1):93–120.

Bailey, J. (1991). Toward a science of metabolic engineering. *Science*, 252(5013):1668–1675.

Bartel, B. and Citovsky, V. (2012). Focus on ubiquitin in plant biology. *Plant Physiology*, 160(1):1–1.

Basler, G., Grimbs, S., Ebenhoh, O., Selbig, J., and Nikoloski, Z. (2012). Evolutionary significance of metabolic network properties. *Journal of the Royal Society Interface,*, 9(71):1168–76.

Becker, S. A. and Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLOS Computational Biology*, 4(5):e1000082.

Becker, S. A., Price, N. D., and Palsson, B. Ø. (2006). Metabolite coupling in genome-scale metabolic networks. *BMC Bioinformatics*, 7(1):111.

Beleggia, R., Rau, D., Laidò, G., Platani, C., Nigro, F., Fragasso, M., De Vita, P., Scossa, F., Fernie, A. R., Nikoloski, Z., and Papa, R. (2016). Evolutionary metabolomics reveals domestication-associated changes in tetraploid wheat kernels. *Molecular Biology and Evolution*, 33(7):1740–1753.

Benedict, M. N., Gonnerman, M. C., Metcalf, W. W., and Price, N. D. (2012). Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon *Methanosarcina acetivorans* C2A. *Journal of Bacteriology*, 194(4):855–865.

Bickel, P. J., Brown, J. B., Huang, H., and Li, Q. (2009). An overview of recent developments in genomics and associated statistical methods. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4313–4337.

Binder, S. (2010). Branched-chain amino acid metabolism in *Arabidopsis thaliana*. *The Arabidopsis book*, 8:e0137.

Borcard, D., Gillet, F., and Legendre, P. (2011). *Numerical Ecology with R*. Springer.

Boulesteix, A.-L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44.

Bradley, P. H., Brauer, M. J., Rabinowitz, J. D., and Troyanskaya, O. G. (2008). Coordinated concentration changes of transcripts and metabolites in *Saccharomyces cerevisiae*. *PLOS Computational Biology*, 5(1):e1000270.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation: Rejoinder. *Journal of the American Statistical Association*, 80(391):614–619.

Burgard, A. P., Nikolaev, E. V., Schilling, C. H., and Maranas, C. D. (2004). Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Research*, 14(2):301–312.

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J. (2007). Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *The Plant Journal*, 52(6):1181–1191.

Çakır, T., Hendriks, M. M. W. B., Westerhuis, J. A., and Smilde, A. K. (2009). Metabolic network discovery through reverse engineering of metabolome data. *Metabolomics*, 5(3):318–329.

Caldana, C., Degenkolbe, T., Cuadros-Inostroza, A., Klie, S., Sulpice, R., Leisse, A., Steinhauser, D., Fernie, A. R., Willmitzer, L., and Hannah, M. A. (2011). High-density kinetic analysis of the metabolomic and transcriptomic response of Arabidopsis to eight environmental conditions. *Plant Journal*, 67(5):869–884.

Campos-de Quiroz, H. (2002). Plant genomics: An overview. *Biological Research*, 35(3-4):385–399.

Cavill, R., Jennen, D., Kleinjans, J., and Briedé, J. J. (2016). Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics*, 17(5):891–901.

Chang, Y., Suthers, P. F., and Maranas, C. D. (2008). Identification of optimal measurement sets for complete flux elucidation in metabolic flux analysis experiments. *Biotechnology and Bioengineering*, 100:1039–49.

Chaves, M. and Sontag, E. D. (2002). State-estimators for chemical reaction networks of Feinberg-Horn-Jackson zero deficiency type. *European Journal of Control*, 8:343–359.

Chen, Y., Zhang, R., Song, Y., He, J., Sun, J., Bai, J., An, Z., Dong, L., Zhan, Q., and Abliz, Z. (2009). RRLC-MS/MS-based metabonomics combined with in-depth analysis of metabolic correlation network: finding potential biomarkers for breast cancer. *Analyst*, 134(10):2003–2011.

Chu, Y. and Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, 22:271–4.

Chuang, H.-Y., Hofree, M., and Ideker, T. (2010). A decade of systems biology. *Annual Review of Cell and Developmental Biology*, 26:721–744.

Chubukov, V., Uhr, M., Le Chat, L., Kleijn, R. J., Jules, M., Link, H., Aymerich, S., Stelling, J., and Sauer, U. (2013). Transcriptional regulation is insufficient to explain substrate-induced flux changes in *Bacillus subtilis*. *Molecular Systems Biology*, 9(1):709.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17:13.

Cosgrove, D. J. (2005). Growth of the plant cell wall. *Nature Reviews Molecular Cell Biology*, 6:850–61.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Dai, Z. and Locasale, J. W. (2017). Understanding metabolism with flux analysis: From theory to application. *Metabolic Engineering*, 43:94–102.

Dall'Osto, L., Holt, N. E., Kaligotla, S., Fuciman, M., Cazzaniga, S., Carbonera, D., Frank, H. A., Alric, J., and Bassi, R. (2012). Zeaxanthin protects plant photosynthesis by modulating chlorophyll triplet yield in specific light-harvesting antenna subunits. *The Journal of Biological Chemistry*, 287:41820–34.

Daran-Lapujade, P., Rossell, S., van Gulik, W. M., Luttik, M. A., de Groot, M. J., Slijper, M., Heck, A. J., Daran, J.-M., de Winde, J. H., Westerhoff, H. V., et al. (2007). The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proceedings of the National Academy of Sciences*, 104(40):15753–15758.

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574.

de Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010a). AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Pshysiology*, 152:579–89.

de Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. (2010b). C4GEM, a genome-scale metabolic model to study C_4 plant metabolism. *Plant Physiology*, 154:1871–85.

Deutscher, J., Francke, C., and Postma, P. W. (2006). How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiology and Molecular Biology Reviews*, 70(4):939–1031.

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, 44(3):362–388.

Donati, S., Sander, T., and Link, H. (2018). Crosstalk between transcription and metabolism: how much enzyme is enough for a cell? *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 10(1).

Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics*, 22(2):101–109.

Du, B., Zielinski, D. C., Kavvas, E. S., Dräger, A., Tan, J., Zhang, Z., Ruggiero, K. E., Arzumanyan, G. A., and Palsson, B. O. (2016). Evaluation of rate law approximations in bottom-up kinetic models of metabolism. *BMC Systems Biology*, 10(1):40.

El-Aneed, A., Cohen, A., and Banoub, J. (2009). Mass spectrometry, review of the basics: electrospray, MALDI, and commonly used mass analyzers. *Applied Spectroscopy Reviews*, 44(3):210–230.

el Bouhaddani, S., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G., and Uh, H.-W. (2016). Evaluation of O2PLS in *omics* data integration. *BMC Bioinformatics*, 17(S2):S11.

Estévez, S. R. and Nikoloski, Z. (2014). Generalized framework for context-specific metabolic model extraction methods. *Frontiers in Plant Science*, 5:491.

Fatland, B. L., Ke, J., Anderson, M. D., Mentzen, W. I., Cui, L. W., Allred, C. C., Johnston, J. L., Nikolau, B. J., and Wurtele, E. S. (2002). Molecular characterization of a heteromeric ATP-citrate lyase that generates cytosolic acetyl-coenzyme A in Arabidopsis. *Plant Physiology*, 130(2):740–56.

Fernie, A. R. and Schauer, N. (2009). Metabolomics-assisted breeding: a viable option for crop improvement? *Trends in Genetics*, 25(1):39–48.

Fernie, A. R. and Tohge, T. (2017). The genetics of plant metabolism. *Annual Review of Genetics*, 51(1):287–310. PMID: 28876980.

Fernie, A. R., Trethewey, R. N., Krotzky, A. J., and Willmitzer, L. (2004). Metabolite profiling: from diagnostics to systems biology. *Nature Reviews Molecular Cell Biology*, 5(9):763–9.

Fiehn, O. (2002). Metabolomics–the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–71.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nature Biotechnology*, 18(11):1157–61.

Fitzgerald, T. L. and McQualter, R. B. (2014). The quantitative real-time polymerase chain reaction for the analysis of plant gene expression. *Cereal Genomics: Methods and Protocols*, pages 97–115.

Floris, M., Mahgoub, H., Lanet, E., Robaglia, C., and Menand, B. (2009). Post-transcriptional regulation of gene expression in plants during abiotic stress. *International Journal of Molecular Sciences*, 10(7):3168–3185.

Friso, G. and van Wijk, K. J. (2015). Posttranslational protein modifications in plant metabolism. *Plant Physiology*, 169(3):1469–1487.

Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B., and Lercher, M. J. (2017). Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLOS Computational Biology*, 13(4):1–14.

Gallant, J. A. (1979). Stringent control in *E. coli*. *Annual Review of Genetics*, 13:393–414.

Gaudinier, A., Tang, M., and Kliebenstein, D. J. (2015). Transcriptional networks governing plant metabolism. *Current Plant Biology*, 3-4:56 – 64.

Gebelein, H. (1941). Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.

Gepts, P. and Papa, R. (2002). *Evolution during Domestication*. In: eLS, John Wiley & Sons, Ltd.

Gerosa, L., van Rijsewijk, B. R. H., Christodoulou, D., Kochanowski, K., Schmidt, T. S., Noor, E., and Sauer, U. (2015). Pseudo-transition analysis identifies the key regulators of dynamic metabolic adaptations from steady-state data. *Cell Systems*, 1(4):270–282.

Giavalisco, P., Hummel, J., Lisec, J., Inostroza, A. C., Catchpole, G., and Willmitzer, L. (2008). High-resolution direct infusion-based mass spectrometry in combination with whole ^{13}C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas. *Analytical Chemistry*, 80(24):9417–9425.

Gibon, Y., Usadel, B., Blaesing, O. E., Kamlage, B., Hoehne, M., Trethewey, R., and Stitt, M. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. *Genome Biology*, 7(8):R76–R76.

Gioia, T., Nagel, K. A., Beleggia, R., Fragasso, M., Ficco, D. B. M., Pieruschka, R., De Vita, P., Fiorani, F., and Papa, R. (2015). Impact of domestication on the phenotypic architecture of durum wheat under contrasting nitrogen fertilization. *Journal of Experimental Botany*, 66(18):5519–30.

Gonçalves, E., Raguz Nakic, Z., Zampieri, M., Wagih, O., Ochoa, D., Sauer, U., Beltrao, P., and Saez-Rodriguez, J. (2017). Systematic analysis of transcriptional and post-transcriptional regulation of metabolism in yeast. *PLOS Computational Biology*, 13(1):1–20.

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., and Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology*, 22:245–52.

Goodey, N. M. and Benkovic, S. J. (2008). Allosteric regulation and catalysis emerge via a common route. *Nature Chemical Biology*, 4:474.

Gu, L., Jones, A. D., and Last, R. L. (2010). Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant. *The Plant Journal*, 61(4):579–90.

Gu, L., Jones, A. D., and Last, R. L. (2012). *Rapid LC-MS/MS Profiling of Protein Amino Acids and Metabolically Related Compounds for Large-Scale Assessment of Metabolic Phenotypes*, pages 1–11. In: Amino Acid Analysis: Methods and Protocols, Humana Press, Totowa, NJ.

Gutiérrez, R. A., MacIntosh, G. C., and Green, P. J. (1999). Current perspectives on mRNA stability in plants: multiple levels and mechanisms of control. *Trends in Plant Science*, 4(11):429–438.

Hackett, S. R., Zanotelli, V. R. T., Xu, W., Goya, J., Park, J. O., Perlman, D. H., Gibney, P. A., Botstein, D., Storey, J. D., and Rabinowitz, J. D. (2016). Systems-level analysis of mechanisms regulating yeast metabolic flux. *Science*, 354(6311).

Hageman Blair, R., Trichler, D. L., and Gaile, D. P. (2012). Mathematical and statistical modeling in cancer systems biology. *Frontiers in Physiology*, 3:227.

Hahn, B. D. (1986). A mathematical model of the calvin cycle: Analysis of the steady state. *Annals of Botany*, 57:639–653.

Hakeem, K. R., Tombuloglu, H., and Tombuloglu, G. (2016). *Plant omics : Trends and Applications*. Springer.

Hannah, M. A., Caldana, C., Steinhauser, D., Balbo, I., Fernie, A. R., and Willmitzer, L. (2010). Combined transcript and metabolite profiling of Arabidopsis grown under widely variant growth conditions facilitates the identification of novel metabolite-mediated regulation of gene expression. *Plant Physiology*, 152(4):2120–2129.

Harrel, F. E. (2015). *Hmisc: Harrell Miscellaneous R package*.

Hartmann, T. (2007). From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry*, 68(22):2831–2846.

Haverkorn van Rijsewijk, B. R. B., Nanchen, A., Nallet, S., Kleijn, R. J., and Sauer, U. (2011). Large-scale (13)c-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*. *Molecular Systems Biology*, 7:477–477.

Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107:1–8.

Heavner, B. D. and Price, N. D. (2015). Transparency in metabolic network reconstruction enables scalable biological discovery. *Current Opinion in Biotechnology*, 34:105–109.

Heinrich, R. and Schuster, S. (1996). The Regulation of Cellular Systems. *Springer*.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28:977.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714.

Horn, F. and Jackson, R. (1972). General mass action kinetics. *Archive for Rational Mechanics and Analysis*, 47(2):81–116.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3-4):321–377.

Hummel, J., Strehmel, N., Bölling, C., Schmidt, S., Walther, D., and Kopka, J. (2013). Mass spectral search and analysis using the golm metabolome database. *The Handbook of Plant Metabolomics*, pages 321–343.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, 8:565–565.

Inouye, M., Kettunen, J., Soininen, P., Silander, K., Ripatti, S., Kumpula, L. S., Hämäläinen, E., Jousilahti, P., Kangas, A. J., Männistö, S., Savolainen, M. J., Jula, A., Leiviskä, J., Palotie, A., Salomaa, V., Perola, M., Ala-Korpela, M., and Peltonen, L. (2010). Metabonomic, transcriptomic, and genomic variation of a population cohort. *Molecular Systems Biology*, 6(1):441.

Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 74(8):2204–2214.

Jain, M. (2012). Next-generation sequencing technologies for gene expression profiling in plants. *Briefings in Functional Genomics*, 11(1):63–70.

Jha, S. and van Schuppen, J. (2001). Modelling and control of cell reaction networks. *Probability Networks and Algorithms*. ISSN:1386-3711.

Johnson, C. H., Ivanisevic, J., and Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7):451–9.

Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4237–4253.

Joloba, M. L., Clemmer, K. M., Sledjeski, D. D., and Rather, P. N. (2004). Activation of the gab operon in an RpoS-dependent manner by mutations that truncate the inner core of lipopolysaccharide in *Escherichia coli*. *Journal of bacteriology*, 186(24):8542–6.

Jorge, T. F., Mata, A. T., and António, C. (2016a). Mass spectrometry as a quantitative tool in plant metabolomics. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 374(2079):20150370.

Jorge, T. F., Rodrigues, J. A., Caldana, C., Schmidt, R., van Dongen, J. T., Thomas-Oates, J., and António, C. (2016b). Mass spectrometry-based plant metabolomics: Metabolite responses to abiotic stress. *Mass Spectrometry Reviews*, 35(5):620–649.

Jozefczuk, S., Klie, S., Catchpole, G., Szymanski, J., Cuadros-Inostroza, A., Steinhauser, D., Selbig, J., and Willmitzer, L. (2010). Metabolomic and transcriptomic stress response of *Escherichia coli*. *Molecular Systems Biology*, 6(1):364.

Kaddurah-Daouk, R., Kristal, B. S., and Weinshilboum, R. M. (2008). Metabolomics: a global biochemical approach to drug response and disease. *Annual Review of Pharmacology and Toxicology*, 48:653–83.

Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J.-I., and Nagata, Y. (2004). Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure*, 12:429–438.

Kaplan, F., Kopka, J., Haskell, D. W., Zhao, W., Schiller, K. C., Gatzke, N., Sung, D. Y., and Guy, C. L. (2004). Exploring the temperature-stress metabolome of Arabidopsis. *Plant Physiology*, 136(4):4159–4168.

Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14:491–496.

Kaufmann, K., Pajoro, A., and Angenent, G. C. (2010). Regulation of transcription in plants: mechanisms controlling developmental switches. *Nature Reviews Genetics*, 11(12):830–842.

Khodayari, A. and Maranas, C. D. (2016). A genome-scale *Escherichia coli* kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nature Communications*, 7:13806.

Khodayari, A., Zomorrodi, A. R., Liao, J. C., and Maranas, C. D. (2014). A kinetic model of *Escherichia coli* core metabolism satisfying multiple sets of mutant flux data. *Metabolic Engineering*, 25:50–62.

Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J., and Lee, S. Y. (2012). Recent advances in reconstruction and applications of genome-scale metabolic models. *Current Opinion in Biotechnology*, 23(4):617 – 623.

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015). Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLOS Computational Biology*, 11(8):e1004321.

Kitano, H. (2007). Towards a theory of biological robustness. *Molecular Systems Biology*, 3:137–137.

Koch, K. E. (1996). Carbohydrate-modulated gene expression in plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, 47(1):509–540. PMID: 15012299.

Kochanowski, K., Gerosa, L., Brunner, S. F., Christodoulou, D., Nikolaev, Y. V., and Sauer, U. (2017). Few regulatory metabolites coordinate expression of central metabolic genes in *Escherichia coli*. *Molecular Systems Biology*, 13(1):903.

Köhler, C. and Springer, N. (2017). Plant epigenomics—deciphering the mechanisms of epigenetic inheritance and plasticity in plants. *Genome Biology*, 18(1):132.

Koshland, D. (1970). The molecular basis for enzyme regulation. In *The Enzymes*, volume 1, chapter 7, pages 341–396.

Kresnowati, M. T. A. P., van Winden, W. A., Almering, M. J. H., ten Pierick, A., Ras, C., Knijnenburg, T. A., Daran-Lapujade, P., Pronk, J. T., Heijnen, J. J., and Daran, J. M. (2006). When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. *Molecular Systems Biology*, 2:49–49.

Krumsiek, J., Bartel, J., and Theis, F. J. (2016). Computational approaches for systems metabolomics. *Current Opinion in Biotechnology*, 39:198–206.

Kurihara, S., Kato, K., Asada, K., Kumagai, H., and Suzuki, H. (2010). A putrescine-inducible pathway comprising puue-ynei in which γ-aminobutyrate is degraded into succinate in *Escherichia coli* K-12. *Journal of Bacteriology*, 192(18):4582–4591.

Ladurner, A. G. (2006). Rheostat control of gene expression by metabolites. *Molecular Cell*, 24(1):1–11.

Lai, Z., Tsugawa, H., Wohlgemuth, G., Mehta, S., Mueller, M., Zheng, Y., Ogiwara, A., Meissen, J., Showalter, M., Takeuchi, K., et al. (2018). Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature Methods*, 15(1):53.

Lal, P. B., Schneider, B. L., Vu, K., and Reitzer, L. (2014). The redundant aminotransferases in lysine and arginine synthesis and the extent of aminotransferase redundancy in *Escherichia coli*. *Molecular Microbiology*, 94(4):843–56.

Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, 44(1/2):289–292.

Ledezma-Tejeida, D., Ishida, C., and Collado-Vides, J. (2017). Genome-wide mapping of transcriptional regulation and metabolism describes information-processing units in *Escherichia coli*. *Frontiers in Microbiology*, 8:1466.

Lei, Z., Huhman, D. V., and Sumner, L. W. (2011). Mass spectrometry strategies in metabolomics. *Journal of Biological Chemistry*, 286(29):25435–25442.

Less, H. and Galili, G. (2008). Principal transcriptional programs regulating plant amino acid metabolism in response to abiotic stresses. *Plant physiology*, 147(1):316–30.

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., and Fernie, A. R. (2006). Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols*, 1(1):387–96.

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. (2011). Controllability of complex networks. *Nature*, 473:167–73.

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. (2013). Observability of complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 110:2460–5.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLOS Computational Biology*, 13(5):1–23.

Lu, P., Rangan, A., Chan, S. Y., Appling, D. R., Hoffman, D. W., and Marcotte, E. M. (2007). Global metabolic changes following loss of a feedback loop reveal dynamic steady states of the yeast metabolome. *Metabolic Engineering*, 9(1):8–20.

Lu, Y., Savage, L. J., Larson, M. D., Wilkerson, C. G., and Last, R. L. (2011). Chloroplast 2010: a database for large-scale phenotypic screening of Arabidopsis mutants. *Plant Physiology*, 155(4):1589–600.

Macrae, R. K. and Long, J. A. (2012). *Transcriptional Regulation in Plants*. John Wiley & Sons, Ltd.

Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9(1):34.

Marashi, S.-A. and Bockmayr, A. (2011). Flux coupling analysis of metabolic networks is sensitive to missing reactions. *Biosystems*, 103(1):57 – 66.

Masuda, S., Mizusawa, K., Narisawa, T., Tozawa, Y., Ohta, H., and Takamiya, K.-I. (2008). The bacterial stringent response, conserved in chloroplasts, controls plant fertilization. *Plant and Cell Physiology*, 49(2):135–41.

Merchante, C., Stepanova, A. N., and Alonso, J. M. (2017). Translation regulation in plants: an interesting past, an exciting present and a promising future. *The Plant Journal*, 90(4):628–653.

Mettler, T., Mühlhaus, T., Hemme, D., Schöttler, M.-A., Rupprecht, J., Idoine, A., Veyel, D., Pal, S. K., Yaneva-Roder, L., Winck, F. V., et al. (2014). Systems analysis of the response of photosynthesis, metabolism, and growth to an increase in irradiance in the photosynthetic model organism chlamydomonas reinhardtii. *The Plant Cell*, 26(6):2310–2350.

Metzner, M., Germer, J., and Hengge, R. (2004). Multiple stress signal integration in the regulation of the complex σ^S-dependent csiD-ygaF-gabDTP operon in *Escherichia coli*. *Molecular Microbiology*, 51(3):799–811.

Meyers, B. C., Galbraith, D. W., Nelson, T., and Agrawal, V. (2004). Methods for transcriptional profiling in plants. Be fruitful and replicate. *Plant Physiology*, 135(2):637–52.

Michaelis, L. and Menten, M. (1913). Die Kinetik der Invertinwirkung. *Biochemische Zeitschrift*, 49:333–369.

Millard, P., Smallbone, K., and Mendes, P. (2017). Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in *Escherichia coli*. *PLOS Computational Biology*, 13(2):e1005396.

Miller, M. B. and Tang, Y.-W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical Microbiology Reviews*, 22(4):611–633.

Mizusawa, K., Masuda, S., and Ohta, H. (2008). Expression profiling of four RelA/SpoT-like proteins, homologues of bacterial stringent factors, in *Arabidopsis thaliana*. *Planta*, 228(4):553–62.

Moco, S., Bino, R. J., Vorst, O., Verhoeven, H. A., de Groot, J., van Beek, T. A., Vervoort, J., and De Vos, C. R. (2006). A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant Physiology*, 141(4):1205–1218.

Mönchgesang, S., Strehmel, N., Trutschel, D., Westphal, L., Neumann, S., and Scheel, D. (2016). Plant-to-plant variability in root metabolite profiles of 19 *Arabidopsis thaliana* accessions is substance-class-dependent. *International Journal of Molecular Sciences*, 17(9):1565.

Moxley, J. F., Jewett, M. C., Antoniewicz, M. R., Villas-Boas, S. G., Alper, H., Wheeler, R. T., Tong, L., Hinnebusch, A. G., Ideker, T., Nielsen, J., and Stephanopoulos, G. (2009). Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. *Proceedings of the National Academy of Sciences*, 106(16):6477–6482.

Nägele, T., Fürtauer, L., Nagler, M., Weiszmann, J., and Weckwerth, W. (2016). A strategy for functional interpretation of metabolomic time series data in context of metabolic network information. *Frontiers in Molecular Biosciences*, 3:6.

Nesbit, M. and Samuel, D. (1998). Wheat domestication: archaeobotanical evidence. *Science*, 279:1431.

Nevoigt, E. (2008). Progress in metabolic engineering of saccharomyces cerevisiae. *Microbiology and molecular biology reviews : MMBR*, 72(3):379–412.

Nguyen, H. V., Müller, E., Vreeken, J., Efros, P., and Böhm, K. (2014). Multivariate maximal correlation analysis. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*.

Nöh, K., Grönke, K., Luo, B., Takors, R., Oldiges, M., and Wiechert, W. (2007). Metabolic flux analysis at ultra short time scale: isotopically non-stationary ^{13}C labeling experiments. *Journal of Biotechnology*, 129:249–67.

O'Brien, E. J., Utrilla, J., and Palsson, B. O. (2016). Quantification and classification of *E. coli* proteome utilization and unused protein costs across environments. *PLOS Computational Biology*, 12(6):e1004998.

Ohama, N., Kusakabe, K., Mizoi, J., Zhao, H., Kidokoro, S., Koizumi, S., Takahashi, F., Ishida, T., Yanagisawa, S., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2016). The transcriptional cascade in the heat stress response of Arabidopsis is strictly regulated at the level of transcription factor expression. *The Plant Cell*, 28(1):181–201.

Oliveira, A. P., Dimopoulos, S., Busetto, A. G., Christen, S., Dechant, R., Falter, L., Haghir Chehreghani, M., Jozefczuk, S., Ludwig, C., Rudroff, F., Schulz, J. C., González, A., Soulard, A., Stracka, D., Aebersold, R., Buhmann, J. M., Hall, M. N., Peter, M., Sauer, U., and Stelling, J. (2015). Inferring causal metabolic signals that regulate the dynamic TORC1-dependent transcriptome. *Molecular Systems Biology*, 11(4):802.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. O. (2011). A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. *Molecular Systems Biology*, 7:535.

Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248.

Pego, J. V., Kortstee, A. J., Huijser, C., and Smeekens, S. C. (2000). Photosynthesis, sugars and the regulation of gene expression. *Journal of Experimental Botany*, 51(90001):407–416.

Peters, S., Janssen, H.-G., and Vivó-Truyols, G. (2010). Trend analysis of time-series data: A novel method for untargeted metabolite discovery. *Analytica Chimica Acta*, 663(1):98–104.

Pilcher, W., Zandkamiri, H., Arceneaux, K., Harrison, S., and Baisakh, N. (2017). Genome-wide microarray analysis leads to identification of genes in response to herbicide, metribuzin in wheat leaves. *PLOS ONE*, 12(12):e0189639.

Postma, P. W., Lengeler, J. W., and Jacobson, G. R. (1993). Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. *Microbiollogical Reviews*, 57(3):543–94.

Price, J., Laxmi, A., St. Martin, S. K., and Jang, J.-C. (2004). Global transcription profiling reveals multiple sugar signal transduction mechanisms in Arabidopsis. *The Plant Cell*, 16(8):2128–2150.

Prost, J. F., Nègre, D., Oudot, C., Murakami, K., Ishihama, A., Cozzone, A. J., and Cortay, J. C. (1999). Cra-dependent transcriptional activation of the icd gene of *Escherichia coli*. *Journal of Bacteriology*, 181(3):893–898.

R-Core-Team (2013). R: A language and environment for statistical computing.

Rahmani, F., Hummel, M., Schuurmans, J., Wiese-Klinkenberg, A., Smeekens, S., and Hanson, J. (2009). Sucrose control of translation mediated by an upstream open reading frame-encoded peptide. *Plant Physiology*, 150(3):1356–1367.

Redestig, H. and Costa, I. G. (2011). Detection and interpretation of metabolite-transcript coresponses using combined profiling data. *Bioinformatics*, 27(13):i357–i365.

Rensink, W. A. and Buell, C. R. (2005). Microarray expression profiling resources for plant genomics. *Trends in Plant Science*, 10(12):603 – 609.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10(3-4):441–451.

Reznik, E., Christodoulou, D., Goldford, J. E., Briars, E., Sauer, U., Segre, D., and Noor, E. (2017). Genome-scale architecture of small molecule regulatory networks and the fundamental trade-off between regulation and enzymatic activity. *Cell Reports*, 20(11):2666–2677.

Rios, D. F., Shirin, A., and Sorrentino, F. (2013). The network observability problem: Detecting nodes and connections and the role of graph symmetries. *arXiv:1308.5261*.

Saha, R., Suthers, P. F., and Maranas, C. D. (2011). Zea mays iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLOS ONE*, 6:e21784.

Salem, M., Bernach, M., Bajdzienko, K., and Giavalisco, P. (2017). A simple fractionated extraction method for the comprehensive analysis of metabolites, lipids, and proteins from a single sample. *Journal of Visualized Experiments : JoVE, (124)*.

Salzman, J., Jiang, H., and Wong, W. H. (2011). Statistical modeling of RNA-seq data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 26(1).

Sanchez, S. and Demain, A. L. (2008). Metabolic regulation and overproduction of primary metabolites. *Microbial biotechnology*, 1(4):283–319.

Savageau, M. A. (1988). Introduction to S-systems and the underlying power-law formalism. *Mathematical and Computer Modelling*, 11:546–551.

Sayikli, C. and Bagci, E. Z. (2011). Limitations of using mass action kinetics method in modeling biochemical systems: illustration for a second order reaction. In *International Conference on Computational Science and Its Applications*, pages 521–526. Springer.

Schallau, K. and Junker, B. H. (2010). Simulating plant metabolic pathways with enzyme-kinetic models. *Plant Physiology*, 152(4):1763–1771.

Schauer, N. and Fernie, A. R. (2006). Plant metabolomics: towards biological function and mechanism. *Trends in Plant Science*, 11(10):508–16.

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., Willmitzer, L., Zamir, D., and Fernie, A. R. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature Biotechnology*, 24(4):447–454.

Schauer, N., Zamir, D., and Fernie, A. R. (2005). Metabolic profiling of leaves and fruit of wild species tomato: a survey of the *Solanum lycopersicum* complex. *Journal of Experimental Botany*, 56(410):297–307. 10.1093/jxb/eri057.

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., and Palsson, B. O. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature Protocols*, 6(9):1290–307.

Schönbrodt, F. D. and Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5):609–612.

Schäuble, S., Stavrum, A. K., Puntervoll, P., Schuster, S., and Heiland, I. (2013). Effect of substrate competition in kinetic models of metabolic networks. *FEBS Letters*, 587(17):2818 – 2824.

Schulze, A. and Downward, J. (2001). Navigating gene expression using microarrays—a technology review. *Nature Cell Biology*, 3(8):E190–E195.

Schwahn, K., Beleggia, R., Omranian, N., and Nikoloski, Z. (2017a). Stoichiometric correlation analysis: principles of metabolic functionality from metabolomics data. *Frontiers in Plant Science*, 8:2152.

Schwahn, K., Küken, A., Kliebenstein, D. J., Fernie, A. R., and Nikoloski, Z. (2016). Observability of plant metabolic networks is reflected in the correlation of metabolic profiles. *Plant Physiology*, 172(2):1324–1333.

Schwahn, K., Omranian, N., and Nikoloski, Z. (2017b). Kschwahn/stoichiometric-correlation v.1.0., doi:10.5281/zenodo.846692.

Schwender, J. and Junker, B. H. (2009). *Plant metabolic networks*. Springer.

Seaver, S. M. D., Gerdes, S., Frelin, O., Lerma-Ortiz, C., Bradbury, L. M. T., Zallot, R., Hasnain, G., Niehaus, T. D., El Yacoubi, B., Pasternak, S., Olson, R., Pusch, G., Overbeek, R., Stevens, R., de Crecy-Lagard, V., Ware, D., Hanson, A. D., and Henry, C. S. (2014). High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. *Proceedings of the National Academy of Sciences*, 111(26):9645–50.

Segal, I. H. (1975). *Enzyme kinetics: Behavior and analysis of rapid equilibrium and steady-state enzyme systems*. John Wiley.

Seo, P. J. and Mas, P. (2014). Multiple layers of posttranslational regulation refine circadian clock activity in Arabidopsis. *The Plant Cell*, 26(1):79–87.

Serin, E. A., Nijveen, H., Hilhorst, H. W., and Ligterink, W. (2016). Learning from co-expression networks: possibilities and challenges. *Frontiers in Plant Science*, 7(44).

Shimizu, K. (2013). Regulation systems of bacteria such as *Escherichia coli* in response to nutrient limitation and environmental stresses. *Metabolites*, 4(1):1–35.

Shulaev, V., Cortes, D., Miller, G., and Mittler, R. (2008). Metabolomics for plant stress response. *Physiologia Plantarum*, 132(2):199–208.

Sims, J. K., Manteiga, S., and Lee, K. (2013). Towards high resolution analysis of metabolic flux in cells and tissues. *Current Opinion in Biotechnology*, 24(5):933–9.

Singh, K. B. (1998). Transcriptional regulation in plants: the importance of combinatorial control. *Plant Physiology*, 118(4):1111–1120.

Singh, V. K. and Ghosh, I. (2006). Kinetic modeling of tricarboxylic acid cycle and glyoxylate bypass in *Mycobacterium tuberculosis*, and its application to assessment of drug targets. *Medical Modelling*, 3:27.

Soong, J. L., Reuss, D., Pinney, C., Boyack, T., Haddix, M. L., Stewart, C. E., and Cotrufo, M. F. (2014). Design and operation of a continuous ^{13}C and ^{15}N labeling chamber for uniform or differential, metabolic and structural, plant isotope labeling. *Journal of visualized experiments: JoVE*, (83).

Sparkman, O. D., Penton, Z., and Kitson, F. G. (2011). *Gas chromatography and mass spectrometry: a practical guide*. Academic Press.

Speranza, M. L., Valentini, G., and Malcovati, M. (1990). Fructose-1,6-bisphosphate-activated pyruvate kinase from *Escherichia coli*. Nature of bonds involved in the allosteric mechanism. *European Journal of Biochemistry*, 191(3):701–704.

Sriyudthsak, K., Shiraishi, F., and Hirai, M. Y. (2016). Mathematical modeling and dynamic simulation of metabolic reaction systems using metabolome time series data. *Frontiers in Molecular Biosciences*, 3:15.

Stanford, N. J., Lubitz, T., Smallbone, K., Klipp, E., Mendes, P., and Liebermeister, W. (2013). Systematic construction of kinetic models from genome-scale metabolic networks. *PLOS ONE*, 8(11):e79195.

Steuer, R., Kurths, J., Fiehn, O., and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19:1019–26.

Stitt, M. (2013). Systems-integration of plant metabolism: means, motive and opportunity. *Current Opinion in Plant Biology*, 16(3):381–388.

Sugliani, M., Abdelkefi, H., Ke, H., Bouveret, E., Robaglia, C., Caffarri, S., and Field, B. (2016). An ancient bacterial signaling pathway regulates chloroplast function to influence growth and development in Arabidopsis. *The Plant Cell*, 28(3):661–679.

Sulpice, R., Nikoloski, Z., Tschoep, H., Antonio, C., Kleessen, S., Larhlimi, A., Selbig, J., Ishihara, H., Gibon, Y., Fernie, A. R., and Stitt, M. (2013). Impact of the carbon and nitrogen supply on relationships and connectivity between metabolism and biomass in a broad panel of arabidopsis accessions. *Plant Physiology*, 162:347–63.

Sulpice, R., Trenkamp, S., Steinfath, M., Usadel, B., Gibon, Y., Witucka-Wall, H., Pyl, E.-T., Tschoep, H., Steinhauser, M. C., Guenther, M., et al. (2010). Network analysis of enzyme activities and metabolite levels and their relationship to biomass in a large panel of Arabidopsis accessions. *The Plant Cell*, 22(8):2872–2893.

Sumner, L. W., Mendes, P., and Dixon, R. A. (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, 62(6):817–836.

Sun, X., Langer, B., and Weckwerth, W. (2015). Challenges of inversely estimating jacobian from metabolomics data. *Frontiers in Bioengineering and Biotechnology*, 3(188):188.

Szymanski, J., Jozefczuk, S., Nikoloski, Z., Selbig, J., Nikiforova, V., Catchpole, G., and Willmitzer, L. (2009). Stability of metabolic correlations under changing environmental conditions in *Escherichia coli* - a systems approach. *PLOS ONE*, 4(10):e7441.

Takahashi, H., Morioka, R., Ito, R., Oshima, T., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Dynamics of time-lagged gene-to-metabolite networks of *Escherichia coli* elucidated by integrative omics approach. *OMICS : a Journal of Integrative Biology*, 15(1-2):15–23.

Tang, Y. J., Martin, H. G., Myers, S., Rodriguez, S., Baidoo, E. E., and Keasling, J. D. (2009). Advances in analysis of microbial metabolic fluxes via ^{13}C isotopic labeling. *Mass Spectrometry Reviews*, 28(2):362–375.

Tchieu, J. H., Norris, V., Edwards, J. S., and Saier, M. H. (2001). The complete phosphotransferase system in *Escherichia coli*. *Journal of Molecular Microbiology and Biotechnology*, 3(3):329–346.

Templeton, G. W. and Moorhead, G. B. (2004). A renaissance of metabolite sensing and signaling: From modular domains to riboswitches. *The Plant Cell*, 16(9):2252–2257.

The MathWorks, I. (2015). *MATLAB 2015a*. The MathWorks, Inc, Massachusetts, United States.

Theodoridis, G., Gika, H. G., and Wilson, I. D. (2011). Mass spectrometry-based holistic analytical approaches for metabolite profiling in systems biology studies. *Mass Spectrometry Reviews*, 30(5):884–906.

Tohge, T., Alseekh, S., and Fernie, A. R. (2014). On the regulation and function of secondary metabolism during fruit development and ripening. *Journal of Experimental Botany*, 65(16):4599–4611.

Tohge, T. and Fernie, A. R. (2009). Web-based resources for mass-spectrometry-based metabolomics: a user's guide. *Phytochemistry*, 70(4):450–456.

Tohge, T. and Fernie, A. R. (2010). Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. *Nature Protocols*, 5(6):1210–1227.

Tohge, T. and Fernie, A. R. (2012). Annotation of plant gene function via combined genomics, metabolomics and informatics. *Journal of Visualized Experiments : JoVE*, (64):3487.

Tohge, T., Scossa, F., and Fernie, A. R. (2015). Integrative approaches to enhance understanding of plant metabolic pathway structure and regulation. *Plant Physiology*, 169(3):1499–1511.

Toubiana, D., Batushansky, A., Tzfadia, O., Scossa, F., Khan, A., Barak, S., Zamir, D., Fernie, A. R., Nikoloski, Z., and Fait, A. (2015). Combined correlation-based network and mQTL analyses efficiently identified loci for branched-chain amino acid, serine to threonine, and proline metabolism in tomato seeds. *The Plant Journal*, 81(1):121–33.

Toubiana, D., Fernie, A. R., Nikoloski, Z., and Fait, A. (2013). Network analysis: tackling complex data to study plant metabolism. *Trends in Biotechnology*, 31(1):29–36.

Toubiana, D., Semel, Y., Tohge, T., Beleggia, R., Cattivelli, L., Rosental, L., Nikoloski, Z., Zamir, D., Fernie, A. R., and Fait, A. (2012). Metabolic profiling of a mapping population exposes new insights in the regulation of seed metabolism and seed, fruit, and plant relations. *PLOS Genetics*, 8(3):e1002612.

Traxler, M. F., Summers, S. M., Nguyen, H.-T., Zacharia, V. M., Hightower, G. A., Smith, J. T., and Conway, T. (2008). The global, ppGpp-mediated stringent response to amino acid starvation in *Escherichia coli*. *Molecular Microbiology*, 68(5):1128–48.

Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128.

Tzin, V. and Galili, G. (2010a). The biosynthetic pathways for shikimate and aromatic amino acids in *Arabidopsis thaliana*. *The Arabidopsis Book / American Society of Plant Biologists*, 8:e0132.

Tzin, V. and Galili, G. (2010b). New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. *Molecular Plant*, 3(6):956–972.

Ulrich, D. and Olbricht, K. (2013). Diversity of volatile patterns in sixteen *Fragaria vesca L. accessions* in comparison to cultivars of *Fragaria xananassa*. *Journal of Applied Botany and Food Quality*, 86(1).

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L., and Fernie, A. R. (2003). Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Reports*, 4(10):989–993.

Ursem, R., Tikunov, Y., Bovy, A., van Berloo, R., and van Eeuwijk, F. (2008). A correlation network approach to metabolic data analysis for tomato fruits. *Euphytica*, 161(1-2):181–193.

Usadel, B. and Fernie, A. R. (2013). The plant transcriptome-from integrating observations to models. *Frontiers in Plant Science*, 4(26):48.

van Kampen, N. G. (2007). *Stochastic Processes in Physics and Chemistry*. North Holland, 3rd edition.

Vance, W., Arkin, A., and Ross, J. (2002). Determination of causal connectivities of species in reaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5816–21.

Veiga, D. F. T., Vicente, F. F. R., Grivet, M., de la Fuente, A., and Vasconcelos, A. T. R. (2007). Genome-wide partial correlation analysis of *Escherichia coli* microarray data. *Genetics and molecular research : GMR*, 6:730–42.

Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J. J., and Yanes, O. (2012). A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites*, 2(4):775–795.

Vlassis, N., Pacheco, M. P., and Sauter, T. (2014). Fast reconstruction of compact context-specific metabolic network models. *PLOS Computational Biology*, 10(1):e1003424.

Voet, D. and Voet, J. G. (2011). *Biochemistry*. Wiley.

Voit, E. O., Martens, H. A., and Omholt, S. W. (2015). 150 years of the mass action law. *PLOS Computational Biology*, 11(1):e1004012.

Wade, J. T. and Struhl, K. (2008). The transition from transcriptional initiation to elongation. *Current opinion in Genetics & Development*, 18(2):130–136.

Wagner, E. M. (2013). Monitoring gene expression: quantitative real-time rt-PCR. *Lipoproteins and Cardiovascular Disease: Methods and Protocols*, pages 19–45.

Walther, D., Strassburg, K., Durek, P., and Kopka, J. (2010). Metabolic pathway relationships revealed by an integrative analysis of the transcriptional and metabolic temperature stress-response dynamics in yeast. *OMICS : a Journal of Integrative Biology*, 14(3):261–274.

Weber, A. P. M., Weber, K. L., Carr, K., Wilkerson, C., and Ohlrogge, J. B. (2007). Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiology*, 144(1):32–42.

Wewer, V., Dombrink, I., vom Dorp, K., and Dörmann, P. (2011). Quantification of sterol lipids in plants by quadrupole time-of-flight mass spectrometry. *Journal of Lipid Research*, 52(5):1039–1054.

Wilson, D. F. (2013). Regulation of cellular metabolism: programming and maintaining metabolic homeostasis. *Journal of Applied Physiology*, 115(11):1583–8.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., et al. (2013). HMDB 3.0—the human metabolome database in 2013. *Nucleic Acids Research*, 41(D1):D801–D807.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52.

Wu, L., Candille, S. I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. *Nature*, 499(7456):79–82.

Yeomans, K. A. and Golder, P. A. (1982). The Guttman-Kaiser criterion as a predictor of the number of common factors. *The Statistician*, pages 221–229.

Yu, X.-Z., Fan, W.-J., Lin, Y.-J., Zhang, F.-F., and Gupta, D. K. (2018). Differential expression of the PAL gene family in rice seedlings exposed to chromium by microarray analysis. *Ecotoxicology*, pages 1–11.

Yugi, K. and Kuroda, S. (2017). Metabolism as a signal generator across trans-omic networks at distinct time scales. *Current Opinion in Systems Biology*, 8:59–66.

Zamboni, N. (2011). ^{13}C metabolic flux analysis in complex systems. *Current Opinion in Biotechnology*, 22(1):103–108.

10 Acknowledgments

A PhD and writing a PhD thesis is something that can not be accomplished alone. Here, I would like to thank and acknowledge all the people, who helped me in the last three years.
First of all, I would like to thank my supervisor Prof. Dr. Zoran Nikoloski for the opportunity to work with him, his supervision and mostly for his patience.

In general, I would like to thank current and former members of the AG Nikoloski, who helped me and answered all my questions or just had a cup of coffee with me. A special thanks to the "coffee people" Michael Scheunemann, Georg Basler and Alain Julio Mbebi. And of curse Jacqueline Nowak for the fruitful discussion about our favorite author: Hildegunst von Mythenmetz.

I would like to thank the International Max Planck Research School "Primary Metabolism and Plant Growth" for funding. A special thanks goes to Ina Talke for her constant support and work as IMPRS coordinator. She always had a time to help with all the problems that can occur during a PhD. Further, I would like to thank all my fellow scholarship holders for the great time during seminars and retreats.

I would like to thank my PhD advisory committee Lothar Willmitzer, Arren Bar-Even, and Marek Mutwill. And I would like to thank Andreas Donath for the technical support.

Further thanks goes to Corné and Yolandi Swart and Semidan Robaina Estévez, for their time reading this thesis and providing helpful comments.

Also, I would like to thank my family and specially to my parents and my fiancée Ricarda for their constant support over all the years. This would not have been possible without them.

11 Statement of authorship

I hereby declare that the present work has not been previously submitted for another degree at any university, has been carried out by myself, and employed only the cited references and resources.

Potsdam, 09.03.2018

Kevin Schwahn