Stoichiometric modelling of plant metabolic networks

Achuthanunni Chokkathukalam Balakrishnan

A thesis submitted in partial fulfilment of the requirements of
Oxford Brookes University
for the award of the degree Doctor of Philosophy

Cell Systems Modelling Group
School of Life Sciences

March 2010

Abstract

Following the completion of the genomic sequencing of the model plant *Arabidopsis thaliana* there has been an increased focus on understanding the characteristics of the interaction between plant metabolism sequestered in various intracellular compartments. A system of such complexity can only be fully encapsulated and understood through the construction of computer models and the use of such models to analyse and interpret experimental data pertaining to the system. This thesis describes the use of steady-state stoichiometric models to study the interaction between the metabolism in chloroplast, cytosol and mitochondria and the application of the characteristics of these models to analyse gene expression data obtained from microarray experiments.

To begin with, independent models of light reactions, glycolysis and the TCA cycle were constructed and a previous model of the Calvin cycle was adapted to suit the purpose of this study. Characteristics of these models with respect to carbon flux were investigated using stoichiometric model analysis techniques such as enzyme subset (ES) analysis, reaction correlation coefficients (RCC) and elementary mode (EM) analysis. The latter identified routes corresponding to the classical metabolic pathways in these compartments and in addition some potential variants.

The independent models were then integrated using relevant transport reactions to study the interactions between chloroplast, cytosol and mitochondria. EM analysis of the integrated model in the absence of net carbon flux revealed a number of routes involved in the exchange of ATP and reducing equivalents generated during light reactions between the three compartments. Previous studies on this topic have demonstrated the role of the triosephosphate-3-phosphoglycerate and the malate-oxaloacetate shuttle mechanisms of the chloroplast membrane in this exchange. The current study exemplifies the existence of other shuttle mechanisms involving glucose-6-phosphate and phosphoenolpyruvate transporters that were not considered earlier. Furthermore, biologically significant modes such as those that may be involved in controlling the over-reduction of chloroplasts were identified.

The subsequent study describes a framework to derive additional information from gene transcript data by structuring it with measures of correlation between reactions derived from metabolic models. The RCCs generated from the integrated model were used as a means to cluster the correlation profiles of genes coding for reactions in the model. The resulting heatmap revealed within- and cross-pathway correlation patterns that may be useful in identifying novel genes and for genome annotation. The heatmap was able to distinguish the compartments in which a particular gene is more highly expressed. This observation was later refined to predict the localisation of enzymes in the model. Furthermore, the heatmap was capable of distinguishing isoforms of genes coding for individual reactions in the model.

To my loving Parents,
any accomplishments of mine are due in no
small part to your support and encouragement.

Acknowledgements

Wow, it's finally time for me to complete this doctoral work! I realise in writing this that I have spent over three years of my life in front of the computer to get to this moment! While I am enthusiastic to finally bring this part of my life to a close, I have to say it has allowed me to get to know, apart from computer modelling and plant physiology, some really outstanding individuals.

I have been most privileged to have worked with undoubtedly the most intuitive, smart and supportive teacher I ever had, namely Prof. David Fell. Ever since I had my first meeting with him, I have been stimulated and excited by his constant flow of good ideas, and his passion for our field of research. He was always there when I needed support — as a supervisor, a good friend and a source of encouragement. He has also known when (and how) to give me a little push in the forward direction when I needed it. Thank you very much for everything you have given me.

I have been indebted in the preparation of this thesis to my co-supervisor, Dr Mark Poolman, whose patience and kindness, as well as his experience in things computational, have been invaluable to me. I thank him for sharing some of his ideas with me, and helping me out with a lot of things that made my life easier.

I must thank my second co-supervisor, Dr Lee Sweetlove, for those invaluable email exchanges, and always guiding me in the right direction when it comes to plant physiology. Some of those articles he send me has had great impact on the outcome of the work described here.

I would like to thank my former colleagues Drs Bushan Bonde, Harshil Patel and Albert Gevorgyan for helping me to take those first steps into doctoral life, many enlightening discussions and constantly letting me know that they had similar problems. Special thanks are due to my colleague, Dr Frances Brightman, for proofreading this thesis amidst her busy schedule. I would like to take this opportunity to thank Miss Chiara Ferrazzi and Dr John Runions for helping me to initiate the ongoing molecular biology work. Thanks to Ms Farida Ben Ghorab, Ms Jill Organ, Mr Philip Voysey and Miss Catherine Hutchinson for helping me with the administrative parts of the course. Special thanks are due to Oxford Brookes University for providing me with the funding for writing this thesis.

One meets many people in graduate school, but some friendships are really special. I want to thank Alexandre Martiniere and his love Maud Vanardois for sharing all those lovely evenings with me. Your friendship has brought a lot of happiness into my life. I would also like to thank my best friend Santhosh Malliot for his 22 year old friendship.

And beyond friends, there is family. My parents Balakrishnan and Sreekumari, and my sister, Poornima, have been there all along. I thank them for supporting me and my

choices throughout my life, and always being there for me. Your sacrifices have made my dreams come true. I also need to thank the family I have inherited through marriage, for showing confidence in me.

Finally, my wife Jesitha Thankappan has been, always, my pillar, my joy and my guiding light, and I thank her. She has seen my best and my worst, and provided support, hugs, and accompanied me to places I never imagined. Even when my emotional and research brains became so hopelessly entwined, she still loved me. And she even thought it was cute.

Thank you all once again.

Contents

List of Abbreviations

Metabolite Abbreviations

2-OG	2-Oxoglutarate
ACoA	Acetyl-CoA
ADP	Adenosine DiPhosphate
AKG	α-Ketoglutarate
AMP	Adenosine MonoPhosphate
ATP	Adenosine TriPhosphate
ASP	Aspartate
BPGA	D-Glycerate 1,3-Bisphosphate
CIT	Citrate
CO_2	Carbon DiOxide
CyB6	Cytochrome b_6f
CytC	Cytochrome C
DHAP	Dihydroxyacetone Phosphate
E4P	D-Erythrose 4-Phosphate
F6P	D-Fructose-6-Phosphate
FAD/$FADH_2$	Flavin Adenine Nucleotide (oxidised/reduced)
FBP	D-Fructose 1,6-Bisphosphate
FD	Ferredoxin
G1P	D-Glucose-1-Phosphate
G6P	D-Glucose-6-Phosphate
GAP	D-Glyceraldehyde 3-Phosphate
GTP	Guanosine TriPhosphate
H_2O	Water
IsoCIT	Isocitrate
MAL	Malate
NAD/NADH	Nicotinamide Adenine Dinucleotide (oxidised/reduced)
NADP/NADPH	Nicotinamide Adenine Dinucleotide Phosphate (oxidised/reduced)
O_2	Oxygen
OOA	Oxaloacetate
PC	Plastocyanin
PEP	Phosphoenolpyruvate
PGA	3-Phospho-D-Glycerate
PGA2	2-Phospho-D-Glycerate
Phe	Pheophytin
P_i	Inorganic Phosphate
PP_i	Inorganic Pyrophosphate
PQ	Plastoquinone

PYR	Pyruvate
Q	Ubiquinone
R5P	D-Ribose 5-Phosphate
RU5P	D-Ribulose 5-Phosphate
RuBP	D-Ribulose 1,5-Bisphosphate
S7P	D-Sedoheptulose 7-Phosphate
SBP	D-Sedoheptulose 1,7-BisPhosphate
SCoA	Succinyl-CoA
SUC	Succinate
UDP	Uridine 5'-DiPhosphate
UDPG	Uridine Diphosphate D-Glucose
X5P	D-Xylulose 5-Phosphate

Reaction Abbreviations

ACN	Aconitate hydratase
AKGDH	Oxoglutarate dehydrogenase (succinyl-transferring)
Ald1	Fructose-bisphosphate aldolase
Ald2	Sedoheptulose-bisphosphate aldolase
CITSynth	Citrate synthase
Complex I	NADH dehydrogenase (ubiquinone)
Complex III	Ubiquinol-cytochrome-c reductase
Complex IV	Cytochrome-c oxidase
FBPase	Fructose-bisphosphatase
FUM	Fumarate hydratase
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GAPDHP	Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)
GlyM	Phosphoglycerate mutase
HK	Hexokinase
IDH	Isocitrate dehydrogenase
NAD-MDH	Malate Dehydrogenase (NAD dependent)
NADP-MDH	Malate Dehydrogenase (NADP dependent)
NADPRe	Ferredoxin-NADP reductase
NDPK	Nucleoside-diphosphate kinase
PEPC	Phosphoenolpyruvate carboxylase
PFK	6-phosphofructo kinase
PFP	Diphosphate-fructose-6-phosphate 1-phosphotransferase
PGI	Phosphoglucose isomerase
PGK	Phosphoglycerate kinase
PGM	Phosphoglycerate mutase
PK	Pyruvate kinase
R5Piso	Phosphoribulose isomerase

Ru5PK	Phosphoribulo kinase
Rubisco	Ribulose-bisphosphate carboxylase
SBPase	Sedoheptulose-bisphosphatase
SCS	Succinate-CoA synthase
SDH	Succinate dehydrogenase
StPase	Starch phosphorylase
StSynth	ADP-glucose starch synthase
SuSyn	Sucrose synthase
TKL1	Transketolase
TKL2	Transketolase
TPI	Triose-phosphate isomerase
TPT_DHAP	DHAP transporter
TPT_PGA	PGA transporter
TPT_GAP	GAP transporter
TX_G6P	G6P transporter
TX_PEP	PEP transporter
TX_PYR	PYR transporter
UGPase	UTP-glucose-1-phosphate uridyltransferase
X5Piso	Phosphoribulose epimerase

Other Abbreviations

ASCII	American Standard Code for Information Interchange
BLAST	Basic Local Alignment Search Tool
BRENDA	Braunschweig Enzyme Database
cDNA	Complementary DNA (single-stranded DNA)
CLI	Command Line Interface
COG	Clusters of Orthologous Groups
COPASI	Complex Pathway Simulator
DNA	Deoxyribonucleic Acid
EM	Elementary Mode
EP	Extreme Pathway
ER	Endoplasmic Reticulum
ES	Enzyme Subset
ETC	Electron Transport Chain
ExPASy	Expert Protein Analysis System
FASTA	Fast All
FBA	Flux Balance Analysis
GFP	Green fluorescent protein
GUI	Graphical User Interface
HGT	Horizontal Gene Transfer
GUI	Graphical User Interface

HGT	Horizontal Gene Transfer
IMS	Intermembrane Space
KEGG	Kyoto Encyclopedia of Genes and Genomes
LP	Linear Programming
MCA	Metabolic Control Analysis
MEGA	Molecular Evolutionary Genetics Analysis
MFA	Metabolic Flux Analysis
MILP	Mixed-Integer Linear Programming
MOMA	Minimisation of Metabolic Adjustment
MS	Mass spectrometry
NASC	European Arabidopsis Stock Centre at Nottingham
NumPy	Numeric Python
ODE	Ordinary Differential Equations
OOP	Object Oriented Programming
OPPP	Oxidative Pentose Phosphate Pathway
PAUP	Phylogenetic Analysis Using Parsimony
PHYLIP	Phylogeny Inference Package
PMF	Proton Motive Force
PSI/PS1	Photosystem I
PSII/PS2	Photosystem II
PySCeS	The Python Simulator for Cellular Systems
RCC	Reaction Correlation Coefficient
RNA	Ribonucleic Acid
SBML	Systems Biology Markup Language
SciPy	Scientific Python
SVD	Singular Value Decomposition
TCA cycle	Tricarboxylic Acid Cycle
T-COFFEE	Tree Based Consistency Objective Function For Alignment Evaluation
URL	Uniform Resource Locator
WPGMA	Weighted Pair Group Method Using Arithmetic Averages

List of URLs

These URLs are indicated by the superscript † throughout this thesis.

Biological databases

AraCyc	http://www.arabidopsis.org/biocyc/index.jsp
BioCyc	http://biocyc.org/
BRENDA	http://www.brenda-enzymes.org/
KEGG	http://www.genome.jp/kegg/kegg1.html
MetaCyc	http://metacyc.org/
NASCArrays	http://affy.arabidopsis.info/narrays/experimentbrowse.pl
SUBA	http://suba.plantenergy.uwa.edu.au/

Software tools

CellNetAnalyzer	http://www.mpi-magdeburg.mpg.de/projects/cna/cna.html
COPASI	http://www.copasi.org/tiki-index.php
FluxAnalyzer	http://www.mpi-magdeburg.mpg.de/projects/fluxanalyzer
GEPASI	http://www.gepasi.org/
iPSORT	http://hc.ims.u-tokyo.ac.jp/iPSORT/
JARNAC	http://www.sys-bio.org/
MEGA	http://www.megasoftware.net/mega.html
METATOOL	http://pinguin.biologie.uni-jena.de/bioinformatik/networks/
MITOPRED	http://bioapps.rit.albany.edu/MITOPRED/
MitoProt II	http://ihg2.helmholtz-muenchen.de/ihg/mitoprot.html
NJPlot	http://pbil.univ-lyon1.fr/software/njplot.html
PAUP	http://paup.csit.fsu.edu/
PeroxiP	http://www.bioinfo.se/PeroxiP/
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
Predotar	http://urgi.versailles.inra.fr/predotar/predotar.html
PyoCyc	http://mudshark.brookes.ac.uk/PyoCyc
PySCeS	http://pysces.sourceforge.net
SCAMP	http://www.sys-bio.org/
ScrumPy	http://mudshark.brookes.ac.uk/ScrumPy
SplitsTree	http://www.splitstree.org
SubLoc	http://www.bioinfo.tsinghua.edu.cn/SubLoc/
TargetP	http://www.cbs.dtu.dk/services/TargetP/
WoLF PSORT	http://wolfpsort.org/
YANA	http://yana.bioapps.biozentrum.uni-wuerzburg.de/

Programming resources

C	http://www.cprogramming.com/
C++	http://www.cprogramming.com/
Java	www.java.com
MATLAB	http://www.mathworks.com/products/matlab/
NumPy	http://numpy.scipy.org
Perl	http://www.perl.org/
Python	http://www.python.org/
SBML	http://sbml.org
SciPy	www.scipy.org
TM4-MeV	http://www.tm4.org/mev/
UML	http://www.agilemodeling.com/artifacts/classDiagram.htm

Preface

This thesis is the final work of my Ph.D. study at the Cell Systems Modelling Group, School of Life Sciences, Oxford Brookes University. It serves as documentation of my work during the study, which has been made from autumn 2006 until spring 2010. The study has been funded by the Oxford Brookes University.

The overall objective of this thesis is to construct and analyse a steady-state stoichiometric model of plant central metabolism. The thesis consists of six chapters. In the first chapter, I have given a general introduction to metabolic modelling and model analysis techniques, and a survey of the recent scientific results. The second chapter provides a general foundation to the aspects of plant metabolism investigated in this thesis. In chapters 3 and 4, I describe the modelling and analysis performed during my study to investigate plant metabolism. In chapter 5, I have presented a framework for integrating metabolic models with gene expression data. The final chapter provides a general discussion on the important outcomes of this thesis and some directions for future work.

Part I

Introduction

CHAPTER 1

Metabolic modelling

1.1 Introduction

Traditional molecular biology research has identified and characterised many of the individual components that make up a living cell and maintain its function. The advent of modern high-throughput molecular biology techniques such as genome sequencing, gene expression profiling, etc. has further accelerated this process. Nonetheless, it is increasingly evident that functions of a biological system can rarely be attributed to an individual component. Instead, it is the interaction amongst the cell's constituents such as proteins, DNA, RNA and small molecules that determine the phenomena observed at higher levels of organisation. A principal objective of modern biology is to describe the structure and dynamics of these intracellular networks of interactions that contribute to the structure and function of a living cell. This perspective and dynamic approach towards understanding the behaviour and properties of biological entities is considered to be in the area of research referred to as *Systems Biology*. Here, the word system is being defined as a set of objects with relations among them, e.g. metabolites and reactions in a metabolic system.

System-level understanding requires a set of principles and methodologies that link the behaviour of individual components to characteristics and functions of the system [1]. In most cases, real-world systems are too complex and understanding them involves identifying the most relevant variables that represent the system and using specific assumptions to represent them. Such an abstract representation of the structure or function of a system that uses mathematical language to describe its behaviour is referred to as a *mathematical model*. A mathematical model uses a set of variables that represent the properties of individual components and a set of equations that establish the relationship between these variables. Mathematical models can thus be used to provide a framework for applying logic and mathematics and for reasoning in a range of situations.

The metabolism in living cells contains a large number of reactions, most of which are capable of converting more than one *substrate* into more than one *product* which inturn is the substrate of one or more reactions. Therefore, a graphical representation in which each substrate is a node and each reaction is an edge will form a metabolic

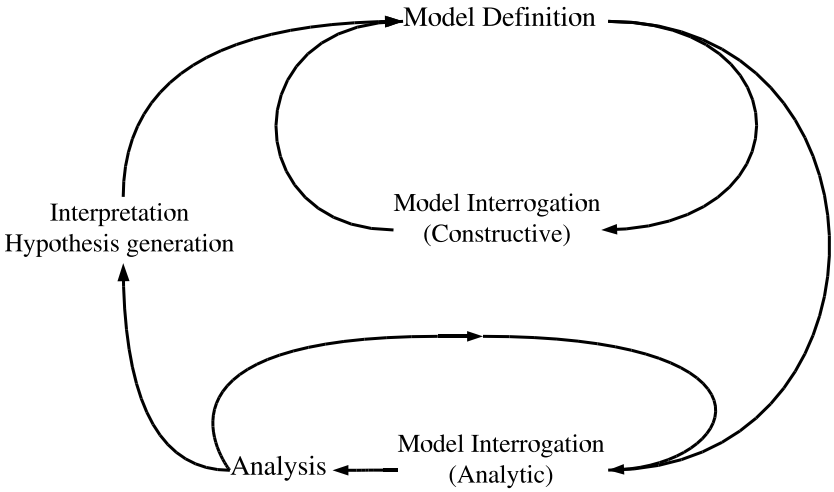


Figure 1.1 – Typical work flow during a modelling investigation. Biologically relevant observations are made during the interpretation/hypothesis generation phase. Adapted from [2].

network. The representation of such a metabolic network in a mathematical form is called a *metabolic model*, where metabolites are the variables and reactions and their kinetics establish the relationship between these variables. One purpose of metabolic modelling is to support experimental design by identifying the variables to measure and the underlying reason for measuring them. Metabolic models can be used for generating hypotheses through prediction and testing these hypotheses. It also tests current knowledge and understanding of the system, i.e. can the known components and their interactions account for observed behaviour. Construction and analysis of metabolic models is, hence, of great scientific, medical and economic importance.

The process of building mathematical models of metabolism is an iterative one. The starting point is an initial hypothesis: that the behaviour of the system under investigation can be explained as a function of the collection of reactions in the model [2]. Models are constructed based on the data collected from literature sources and on-line databases such as KEGG† [3], MetaCyc† [4] and BRENDA† [5] (note that † superscript will be used throughout this thesis to direct the reader to the List of URLs section). Once the model is defined, it is interrogated to extract valuable information. This stage of model building can be divided into two distinct phases: constructive and analytic. Due to reasons that range from a trivial mistyping of some reaction information to something more serious such as the omission of an essential reaction, the initial versions of the model usually have very low interpretive and predictive capabilities. During constructive interrogation such errors in the model are identified and rectified and the process is repeated iteratively, until a sufficient level of accuracy is achieved. In the next step, analytic interrogation, the model is used to study the underlying biological

properties and behaviour of the system. The results of such an interrogation may either lead to further refinement of the model or generate further hypotheses [2] (Figure 1.1).

A number of different formalisms are employed in constructing metabolic models, the most common being kinetic and structural models. A kinetic model represents a quantitative approach to metabolic modelling and contains kinetic information that defines the temporal behaviour of a system, starting from a given initial state. For this reason, it can be used to describe the time-dependent changes of the variables of a system (e.g. metabolite concentrations, reaction rates) when the experimental system is perturbed [6]. Kinetic modelling has been successfully applied to small models in which kinetic parameters are already available or are possible to quantify. However, large kinetic models are too complex to design and implement because of the huge volume of quantitative information required (which is often not available) and the limitations in the currently available theoretical and computational capabilities [7, 8]. For similar reasons, the results obtained from such an analysis would be difficult to interpret and may not reflect the original characteristics of the system under study.

Structural models on the other hand leave aside the many kinetic information and consider some basic constraints arising from the network structure and thermodynamic principles. These models are described purely in terms of the stoichiometries[1] of reactions in the system, which are often readily available. The exclusion of kinetic data restricts structural models in terms of the level of predictions that can be made for a given system. Nevertheless, this is in some ways compensated by the ability to build larger structural models and determine a variety of model properties that could not be found by any other means.

Models are either hand-built or automatically generated. Hand-built models are a closer reflection of the system under study as the size of the system permits clear and precise constructive and analytic interrogation of the entities involved in the model. They are very useful for studying small structural networks (e.g. glycolysis) or for kinetic investigations where the reaction parameters are already known. An advantage with stoichiometric models is that they are based on well-known stoichiometric coefficients and that they do not require determination of kinetic parameters. With the increasing amount of genome-sequencing and annotation efforts being undertaken, it is therefore relatively straightforward to automatically construct organism-specific stoichiometric models of metabolism. In recent years such large-scale models, also called genome-scale models, primarily based on genome sequence information have been developed. The modeled organisms include many prokaryotes such as *Haemophilus influenzae* [9], *Escherichia coli* [10] and *Helicobactor pylori* [11], and eukaryotes such as *Saccharomyces cerevisiae* [12] and most recently the model

[1] The stoichiometry of a reaction expresses the quantitative relationship between reactants and products in a chemical reaction in mole numbers. See Section 1.3.1 for an example.

plant *Arabidopsis thaliana* [13]. They have been used for computational studies and for predicting the network properties and cellular behaviour under different physiological conditions. However, when building models of this size (i.e. typically in excess of 250 reactions) the precision attributed to small hand-built models is diminished.

For all these reasons, the modelling and analysis described in this thesis solely employ structural modelling techniques. In subsequent sections of this chapter I describe the mathematical foundations of structural modelling, various methods involved in analysing structural models and the software employed in modelling metabolism.

1.2 Structure of a metabolic model

A structural metabolic model typically consists of a selected list of metabolic reactions and the metabolites they are associated with, along with a description of the environment within which the system resides. In order to construct a sufficiently realistic model that reflects the *in vivo* characteristics of the system under interrogation, certain frameworks apply when defining these entities in a metabolic model.

1.2.1 Metabolites

The set of metabolites in a metabolic model is subdivided to two subsets: *internal* and *external* metabolites, depending on their relation to the model's boundary. Internal metabolites are defined within the model as those that are likely to be produced and consumed as part of the intracellular metabolism of the system under consideration.

On the other hand, external metabolites are assumed to have concentrations that are maintained by the environment or large enough that the changes caused by the reaction system become negligible. Characteristics that determine the externality of a metabolite are as follows:

- Source or sink metabolites that are consumed (e.g. glucose) and/or produced (e.g. ethanol) by the system. Their amount is usually assumed to be constant, due to availability in large excess or well-tuned biological regulation.
- Metabolites that are in constant exchange with the extracellular environment, such as water and carbon dioxide, whose concentrations are not affected by the reactions in the model.
- Any polymeric metabolite, such as proteins (e.g. myoglobin and albumin), nucleotides (e.g. DNA and RNA) and polysaccharides (e.g. starch and glycogen), whose stoichiometry does not imply the number of monomers incorporated into it.

- Metabolites that are highly connected (i.e. participate in many reactions) within the model (e.g. protons, ATP/ADP and NAD^+/NADH). They may be made external to reduce the connectivity within the model and to increase the interpretive and predictive capabilities of the model.

In some cases, externality of a particular metabolite is determined by the expected outcome of the modelling investigation. For example, by changing the source or sink metabolites the behaviour of the system under different environmental conditions can be studied.

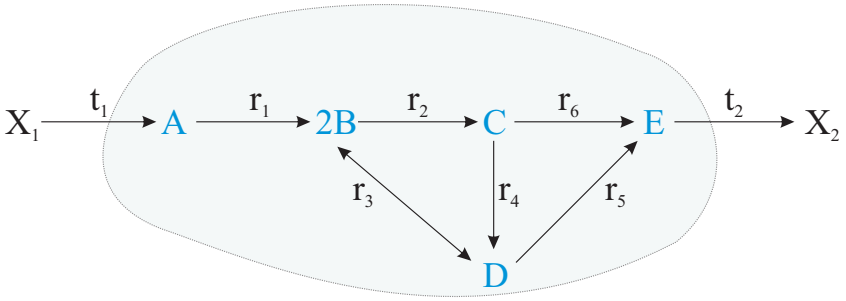
1.2.2 Reactions

A reaction is the conversion of one set of metabolites, called substrates, into another set, called products (Figure 1.2(a)), in amounts specified by the reaction stoichiometry. Most reactions are catalysed by enzymes, while the others are 'spontaneous'. The set of reactions in a metabolic model can be subdivided into subsets of reversible (bi-directional) and irreversible (uni-directional) reactions. In principle, all biochemical reactions are reversible. However, some reactions can be considered irreversible *in vivo* if they exclusively proceed in one direction. In a metabolic model, reactions are defined as irreversible in order to maintain the metabolic flux and to determine the directionality of the overall process. Reversibility of a reaction is an important aspect to be considered while defining a metabolic model [14].

An exception to normal reactions are transport reactions, which do not necessarily involve any enzymatic conversions where the substrate and the product are the same chemical substance. Transport reactions are mediated by a set of molecules called transporters, which account for the movement of metabolites between cellular compartments. In structural models, such reactions are represented as interconversions of metabolites which refer to the same chemical species. For example, import of glucose into the cell can be represented as x_glucose $\rightarrow$ glucose, where the substrate and product represent the external and internal 'versions' of glucose, respectively. Those reactions that consume or produce external metabolites are referred to as 'exchange reactions'.

1.3 The mathematical basis of metabolic modelling

There exist a number of fundamental concepts and principles based on which metabolic models can be successfully defined and analysed. The following sections will aim to describe some of these concepts with appropriate examples, to a level that is required for understanding the approaches taken in this thesis.



(a)

t_1 : $X_1 \longrightarrow A$
r_1 : $A \longrightarrow B$
r_2 : $2B \longrightarrow C$
r_3 : $2B \longleftrightarrow D$
r_4 : $C \longrightarrow D$
r_5 : $D \longrightarrow E$
r_6 : $C \longrightarrow E$
t_2 : $E \longrightarrow X_2$

(b)

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccccccc} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 & t_1 & t_2 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & -2 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & -1 \end{bmatrix} \end{array}$$

(c)

Figure 1.2 – Formalisms in representing metabolic models. (a) A simple metabolic model. X_1 and X_2 are external metabolites whose concentrations are fixed. t_1 and t_2 represent the exchange reactions that consume and produce external metabolites, respectively. A, B, C, D and E are internal metabolites whose rates of formation equal their rates of utilisation. r_1, r_2, r_3, r_4, r_5 and r_6 are reactions. Arrows show the direction of flow of matter. (b) A symbolic list of reactions representing our simple metabolic model. (c) Stoichiometry matrix **N** representing the reaction pathway specified in Figure 1.2(a). Columns and rows correspond to reactions and metabolites involved in the reaction, respectively. The elements of the matrix represent stoichiometric coefficients of the metabolites in the corresponding reaction and their symbol denotes whether they are consumed (-) or produced (+).

1.3.1 Representation of metabolic models

The interconnectivity of metabolites within a network of biochemical reactions is represented by reaction equations defining the stoichiometric conversion of substrates into products for every reaction. A number of distinct formalisms exist for the representation of such biochemical reaction networks. Figure 1.2(a) illustrates the representation of a metabolic model as seen in basic chemistry textbooks and the symbolic list of reactions shown in Figure 1.2(b) is widely used as input in modelling software and for storage in databases. However, all methods of analysis require a mathematical representation of the metabolic model, the starting point for which is provided by the stoichiometry matrix (Figure 1.2(c)) [15]. The stoichiometry matrix contains all the information about how substances are linked through reactions within the network. It indicates the topological structure and architecture of the network, and a knowledge of its properties is prerequisite for any mathematical analyses of biological reaction networks [16, 17].

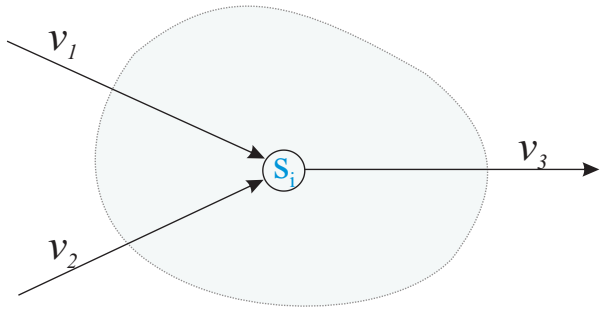


Figure 1.3 – An illustration to demonstrate the steady-state approximation.

Coefficients of all reactions in a system can be used to create a stoichiometry matrix **N** of dimension $m \times n$, whose rows m and columns n correspond to the internal metabolites and reactions, respectively. Note that the stoichiometry matrix representing our simple metabolic model (Figure 1.2(c)) only contains rows representing the internal metabolites and columns representing the internal reactions (including exchange reactions). Such a matrix is called the *internal* stoichiometry matrix[2]. Each element n_{ij} in this matrix represents the stoichiometric coefficient of metabolite i in reaction j. '-' and '+' signs are used to denote whether a particular metabolite is consumed or produced. '0' represents a metabolite that is neither consumed nor produced. Construction of an internal stoichiometry matrix is the extent of the system definition required for many modes of model interrogation. Analytical methods based on the direct interrogation of **N** can be employed to study a number of properties of the metabolic network and are described further in Section 1.4.1.

In cases where the production and consumption of the external metabolites are to be computed, an *external* stoichiometry matrix, with additional rows representing the involvement of externals in the reaction, has to be constructed [15]. The external stoichiometry matrix is particularly useful in analysing mass flow through the system.

1.3.2 The steady-state approximation

One characteristic feature of biological systems is that they are open, that is, they interact with their environment through the exchange of matter and energy. This means that there is a constant flux of source and sink metabolites into and out of the system, respectively. However, in case of a non-growing system, the total mass remains conserved as the net import of mass into the system per unit time is equal to the net export per unit time. This situation of the system is referred to as the *steady state* [18, 19].

[2] Unless stated otherwise **N** represents the internal stoichiometry matrix in this thesis.

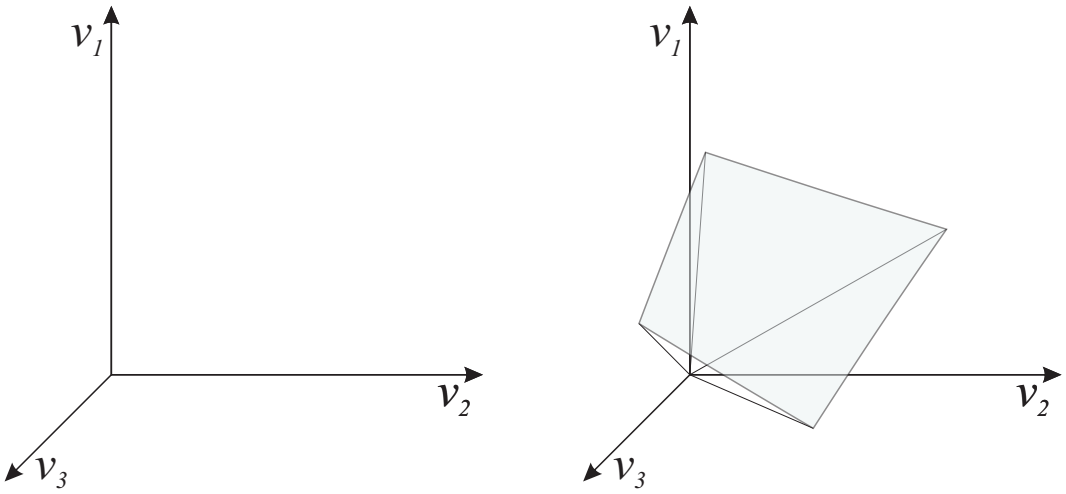


Figure 1.4 – Graphical representation of the subspace. (a) A three dimensional space where the axes represent fluxes through all individual reactions in the metabolic network. (b) The steady-state constraint imposed on the stoichiometry matrix limits the fluxes to a subspace. Adapted from [8].

The steady-state approximation is extensively used in metabolic modelling to study complex biochemical systems as it allows the modeller to define the concentrations of all pathway substrates and products as constant. This implies that the rates of formation of all internal metabolites must always be equal to their rates of utilisation and that their total concentration is time invariant [19]. Consider the simple metabolic system shown in Figure 1.3: if **S** is the concentration of internal metabolites then total input into the system must equal the net output from the system in order to keep **S** constant. Under the steady-state approximation, the rate of change of concentration of a single metabolite s_i is given by:

$$\frac{ds_i}{dt} = v_1 + v_2 - v_3 = 0. \tag{1.1}$$

In general, the rate of change of the concentration of a metabolite is the sum of the reaction rates, each weighted by the corresponding stoichiometric coefficient of the metabolite. If **v** and **s** represent vectors whose elements correspond to the reaction rates and the concentration of metabolites, respectively, in a metabolic system, then the rate of change of the concentration of all metabolites in the system can concisely be written as:

$$\frac{d\mathbf{s}}{dt} = \mathbf{N} \cdot \mathbf{v}. \tag{1.2}$$

At steady state, these concentrations are constant:

$$\mathbf{N} \cdot \mathbf{v} = 0. \tag{1.3}$$

As an example, our steady-state model shown in Figure 1.3 gives the equation

$$\frac{d\mathbf{s}}{dt} = \begin{bmatrix} 1 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0. \tag{1.4}$$

Both kinetic and structural modelling are based on the stoichiometric or mass balance constraint imposed on the system by Equation 1.3. The equation lends itself to further analysis centred around the concepts of linear algebra and allows us to employ mathematical techniques and concepts to solve biological problems. In kinetic modelling **v** are functions of **s** that are solved to determine the concentrations of the substrates at steady state.

1.3.3 The null-space matrix

In structural modelling, fluxes (**v**) in the homogeneous system[3] of linear equations represented by Equation 1.3 are considered as variables and hence, the equation is usually underdetermined as there are infinitely many flux solutions. The trivial solution $\mathbf{v} = 0$ always fulfils Equation 1.3. However, this would mean that all reactions in the system do not carry flux. Non-trivial and biologically meaningful solutions for Equation 1.3 can be obtained using linear algebraic methods by calculating the subspace of all possible solutions. This subspace is called the null space and it can be derived from the stoichiometry matrix **N** either by applying Gaussian elimination [20, 21] or Singular Value Decomposition (SVD) [22, 23] to it. Figure 1.4(b) shows a graphical representation of all solutions of the null space obtained from the stoichiometry matrix, where every point within the subspace represent a solution (flux distribution) that obeys the steady-state assumption.

A null space can be described mathematically by a *kernel* matrix **K**[4], whose columns are linearly independent[5] vectors that together form a *basis* spanning the vector space [21, 24]. For example, if the set of vectors $\mathbf{B} = \{v_1, v_2, v_3, ..., v_n\}$ are linearly independent and span the vector space **K**, then the set **B** is called the basis of the vector space **K**. The null space of the stoichiometry matrix **N** is the set of all vectors in **K** such that:

$$\mathbf{N}.\mathbf{K} = 0. \tag{1.5}$$

It is often easiest to describe the null space of a matrix by finding a basis for the null space. The number of vectors in the basis is called the *dimension* of the null space. In a

[3] A linear equation $f(x) = C$ is called homogeneous if $C = 0$.

[4] Unless stated otherwise, **K** will represent the (right) null-space of **N** henceforth in the thesis.

[5] A vector **b** is called a linear combination of the vectors $v_1, v_2, v_3, ..., v_n$ if it can be expressed in the form $\mathbf{b} = x_1v_1 + x_2v_2 + x_3v_3 + ... + x_nv_n$, where $x_1, x_2, x_3, ..., x_n$ are scalars. None of the vectors in **K** can be written as a linear combination of finitely many other vectors in the collection.

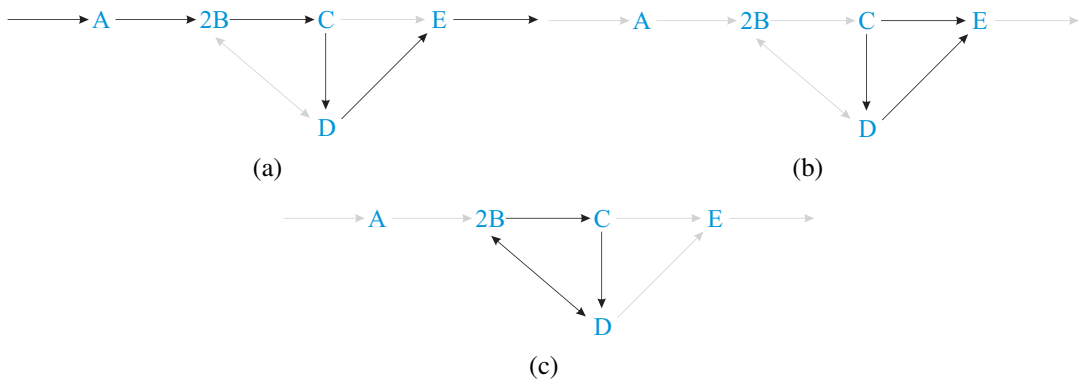


Figure 1.5 – Flux distributions (black arrows) representing the null-space vectors (Equation 1.6) of the simple metabolic model in Figure 1.2(a). The external metabolites are not shown.

metabolic model, the dimension of the null-space is the difference between the number of reactions in the model and the rank of the stoichiometry matrix **N** (rank(**N**)). The null space contains all of the possible solutions and hence flux distribution vectors (v) that satisfy Equation 1.3.

However, certain limitations exist in the representation and analysis of the null space represented by the kernel matrix. As an example, consider our simple metabolic network in Figure 1.2(a). A possible basis for the null space of the stoichiometry matrix (Figure 1.2(c)) is:

$$\mathbf{K} = \begin{bmatrix} -2 & 0 & 0 \\ -2 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \tag{1.6}$$

where the three vectors correspond to the flux solutions a, b and c shown in Figure 1.5. In biochemical terms, negative values for irreversible reaction rates are impossible and therefore vectors with negative values have no biochemically meaningful information [16]. The basis vectors in Equation 1.6 and its representation in Figure 1.5 indicate that the kernel matrix, in general, is not unique. For example, one of the columns of **K** can be substituted with another feasible flux distribution vector (e.g. $\mathbf{v} = (1,1,1,0,0,0,1,1)^T$). Hence, the real number of the feasible metabolic fluxes in the model can only roughly be estimated from the dimensions of **K** [25]. Note also that the kernel matrix does not take into consideration either the reversibility or the capacity restrictions of the reactions in the model [25]. Some of these shortcomings of the null space are overcome by constraint-based approaches described in Section 1.4. Nevertheless, some very important steady-state properties of metabolic models can be derived from the null space.

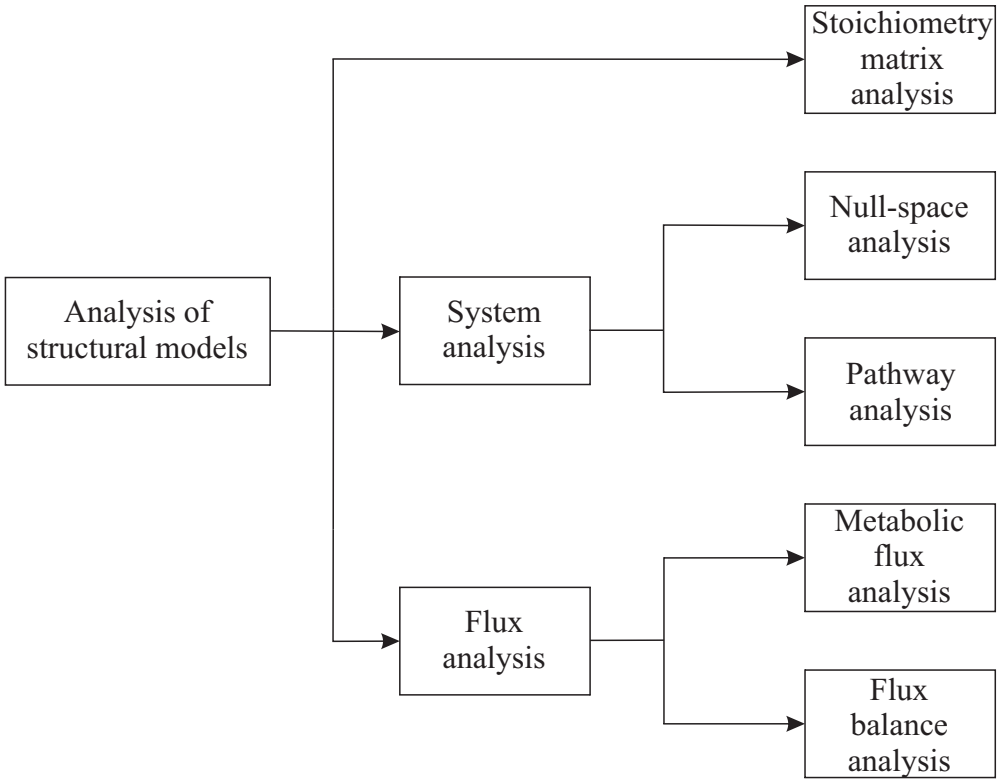


Figure 1.6 – Methodologies and approaches employed in analysing structural metabolic models. Adapted in part from [27].

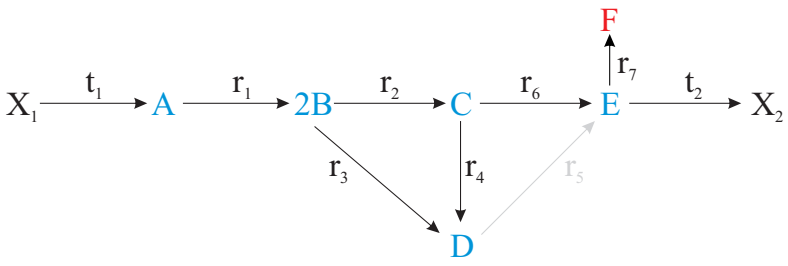
1.4 Analysing structural models

A variety of approaches are employed in analysing structural metabolic models. The majority of the approaches depend on certain constraints that restrict the behaviour of the model, such as the steady-state mass balance of metabolites or the irreversibility of reactions due to thermodynamic limitations (Table 1.1). These constraints can be both invariant (i.e. non-adjustable) or adjustable [8]. The former can be used to analyse the general solution space that encompasses all possible steady-state behaviours (e.g. flux distributions) of the system and the latter to identify particular behaviours within the allowable solution space, such as behaviours that produce the highest possible growth rate or production of a particular metabolite [26].

Approaches to analysing structural models can, hence, be rationally classified based on the constraints employed as those that focus on the properties of the entire space

Table 1.1 – Physiochemical constraints used for analysing structural metabolic models. Adapted from [8] and [27].

Constraint	Type	Mathematical formulation
Systemic stoichiometry	Invariant	$\mathbf{N}.\mathbf{v} = 0$ (defines the solution space)
Irreversibility of reactions	Invariant	$v > 0$
Enzyme/transporters capacities	Invariant	$v < v_{max}$
Measured fluxes	Adjustable	$v = v_m$ or $v_{m,min} < v < v_{m,max}$
Regulatory constraints	Adjustable	Example: $v_1 = 0$ if $(v_2 \neq 0)$



(a)

	r_1	r_2	r_3	r_4	r_6	r_7	t_1	t_2	
A	−1	0	0	0	0	0	1	0	
B	1	−2	−2	0	0	0	0	0	
C	0	1	0	−1	−1	0	0	0	
D	0	0	1	1	0	0	0	0	←
E	0	0	0	0	1	−1	0	−1	
F	0	0	0	0	0	1	0	0	←

(b)

Figure 1.7 – Orphan and dead-end metabolites. (a) The simple metabolic model in Figure 1.2(a) is slightly modified to indicate an orphan (F) and a dead-end (D) metabolite. Reaction r_3 or r_4 can be made reversible to make D live. (b) Presence of these metabolites in a model can be identified from its stoichiometry. Notice the stoichiometry of the metabolites D and E.

of possible flux distributions, and those for determining particular flux solutions in it. Another very useful approach to identify the underlying properties of a metabolic network is the direct investigation of its stoichiometry matrix. An illustration of the major approaches involved in analysing metabolic models can be seen in Figure 1.6. The rest of this section has been organised with the objective of a convenient exposition of the major approaches, rather than as a full-blown review. Wherever possible, readers are directed to detailed reviews on selected topics.

1.4.1 Stoichiometry matrix analysis

Stoichiometry defines the relationship between reactants and products in a balanced chemical reaction. The stoichiometry matrix forms the most basic feature of a biochemical reaction network. Direct analysis of the stoichiometry matrix enables us to draw useful conclusions regarding the inherent network structure and the organisation of the metabolites and reactions in the network.

1.4.1.1 Orphan and dead-end metabolites

Despite the care and effort with which a structural model may be constructed, the resulting network can fall short of biological expectation, the most likely problem being the existence of a large number of *orphan* or *dead-end* metabolites. The former are

those metabolites in the model that are either produced and consumed by a single reaction. Metabolites of this type clearly cannot be balanced, and hence reactions involved with them, and quite possibly additional reactions, must be dead[6] [28]. Dead-end metabolites, on the other hand, are involved in more than one reaction, but are neither produced nor consumed by another reaction. Like orphan metabolites they will also result in dead reactions, but this problem can be resolved by making one of the reactions involved with them reversible.

The presence of any of these metabolites will indicate the existence of disconnected subnetworks within the model. In models that are generated directly from databases, both kind of metabolites can originate either from inaccuracies or inconsistencies in the database, or from incorrect interpretation of the data contained therein [28]; in other cases the situation simply reflects a lack of knowledge.

Orphan and dead-end metabolites in a metabolic model can be identified from its stoichiometry matrix. While rows with only one negative or positive coefficient indicate orphan metabolites, rows with only negative or positive coefficients represent dead-end metabolites. Identification and effective handling of these metabolites is a major step in model definition and constructive interrogation.

1.4.1.2 Conservation relations and conserved moieties

A characteristic feature of biological networks is the conservation of certain molecular subgroups, termed moieties [16, 29]. A typical example of a conserved group is the conservation of the adenine nucleotide moiety, i.e. the total amount of ATP, ADP and AMP is constant during the evolution of the system. Other common examples include the conservation of pyridine nucleotides between NAD^+ and NADH, proteins between phosphorylated and unphosphorylated states, and so on. When one of these cosubstrates is consumed, the other is produced, keeping the sum of both concentrations constant. Figure 1.8(a) illustrates a simple network that displays a conserved moiety, in this case the total mass, M1 + M2 remains constant during the evolution of the network.

A general property of any conservation relation τ between moieties in a metabolic model is that it represents a combination of rows in its stoichiometry matrix that are linearly dependent[7] [16]. Linear combination of the rows of the stoichiometry matrix $\mathbf{N}$ can be represented by $\mathbf{N}^T.\tau$, where $\mathbf{N}^T$ is the transpose of $\mathbf{N}$. To find linearly dependent rows τ must fulfil:

$$\mathbf{N}^T.\tau = \mathbf{0}, \tag{1.7}$$

where $\mathbf{0}$ is a zero column ($m \times 1$) vector. This means that τ must lie in the null space of the transpose of $\mathbf{N}$ (also called the left null space of $\mathbf{N}$ [22]). The dimension of the

[6] Reactions that cannot carry flux at steady-state (Section 1.4.2.1).

[7] See footnote on page 10 for more on linear dependency in a matrix.

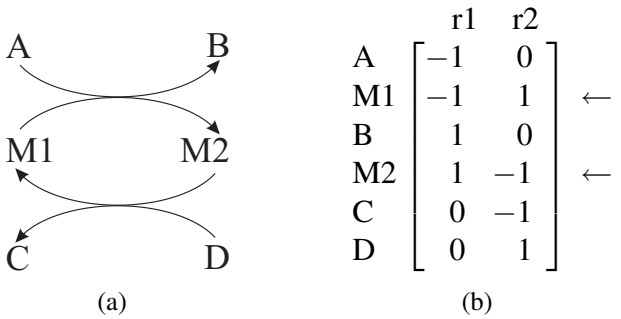


Figure 1.8 – A simple conservation relationship model (a) and its stoichiometry matrix (b). Notice the two rows representing the stoichiometry of M1 and M2 ([-1,1] and [1,-1] respectively). Since either row can be derived from the other by multiplication by -1, they are linearly dependent.

left null space represent the number of linearly independent conservation relations in a model and can be obtained by subtracting the number of rows of **N** (m) with the rank of **N** (rank(**N**)) [29]. Identifying conservation relations is useful in detecting conserved moieties in metabolic models.

Interested readers are referred to Sauro and Ingalls [29] for a complete description of several methods and corresponding algorithms for the determination of conservation relationships in a metabolic model.

1.4.1.3 Graph-based methods

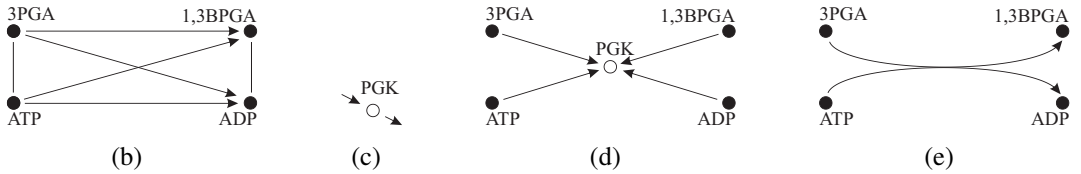
The biochemical interactions in biological networks can be conveniently represented as mathematical graphs, in which the nodes (also called vertices) represent the constituent building blocks (e.g., genes, proteins, metabolites, etc.), and the edges (links connecting pairs of vertices) represent the interactions between them [30, 31]. Depending on the nature of these interactions, edges can be directed or non-directed. In graphs with directed edges, the interaction between any two nodes has a well-defined direction, which represents, for example, the flux from a substrate to a product in a metabolic reaction, whereas graphs with non-directed edges are used to represent mutual interactions as their edges do not have an assigned direction [31].

The mathematical notation for a graph composed of N nodes and E edges is $G(N,E)$. The structure of this graph can be represented by means of an *adjacency matrix* **A(n,m)** whose element a_{ij} is 0 if the nodes are not connected and 1 otherwise. Graphs representing metabolic networks can be constructed from the stoichiometry matrix by converting it to an appropriate adjacency matrix [32].

A number of formalisms exist for the representation of metabolic networks as graphs (Figure 1.9). In substrate graphs, the metabolites are represented as nodes and connected by edges, if they occur in the same reaction. Whereas in reaction graphs, reactions are represented as nodes and are interconnected if they use at least one common metabolite. A bipartite graph is characterised by two separate classes of

3-phosphoglycerate (3PGA) + ATP $\xrightarrow{PGK}$ 1,3 bisphosphoglycerate (1,3BPGA) + ADP

(a)



(b) (c) (d) (e)

Figure 1.9 – A simple reaction catalysed by the enzyme phosphoglycerate kinase (PGK) (a) is represented using the four formalisms in representing metabolic graphs, namely, substrate graphs (b), reaction graphs (c), bipartite graphs (d) and hypergraphs (e).

nodes representing reactions and metabolites, the edges connecting the reactions to the metabolites that they interconvert. In this type of graph edges are used to represent additional data (called edge weights) such as stoichiometric coefficients, confidence levels, strengths, or reaction speeds. Examples for the bipartite graph representation of metabolic models can be found in the KEGG database. Another familiar method for representing metabolic graphs is called hypergraphs, where an edge relates a set of substrates to a set of products. This type of graph is found in MetaCyc.

The representation of complex biochemical networks as graphs has made it possible to systematically investigate the topology and function of these networks using well-defined graph theoretical concepts. Numerous measures have been defined for this purpose; they include:

- *Degree*

  The nodes of a graph can be characterised by the number of edges that they have or the number of other nodes to which they are adjacent [31]. This property is called the node degree or connectivity, k. In directed graphs, there is an incoming degree (in-degree) which denotes the number of edges that point to a node, and an outgoing degree (out-degree) which denotes the number of edges that originate from it.

- *Degree distribution*

  While node degrees characterise individual nodes, the degree distribution can be used to quantify the diversity of the whole graph. The degree distribution $P(k)$ can be calculated by counting the number of nodes $N(k)$ that have k = 1, 2, 3... edges and dividing it by the total number of nodes N [30]. The degree distributions of metabolic networks have been claimed to follow a well-defined functional form $P(k) \sim Ak^{\gamma}$, called the power law, where γ is the degree exponent and $\sim$ indicates 'proportional to' [33, 34]. The value of γ determines the role of highly connected

nodes[8] in the graph and is usually in the range $2 < \gamma < 3$ [35]. The absence of a typical degree is why these networks are called *scale free* [31].

- *Shortest path and mean path lengths*

  The number of edges that exist between two nodes on a graph is called the path and the shortest path connecting these two nodes is called the path length. In most graphs, there is a relatively short path between any two nodes, and its length is in the order of the logarithm of the network size [35]. This small-world property is claimed by some to characterise most complex networks, including metabolic networks [34].

Measures like these can be used to predict the structural and dynamic properties of the underlying network. Such predictions can suggest new biological hypotheses and drive subsequent experimentation. Furthermore, novel graph analysis techniques employing these measures may provide powerful tools to address fundamental biological questions at the system level. One such technique referred to as *damage analysis*, is used to investigate the extent of loss induced in a metabolic graph by the removal of a single enzyme [36]. The analysis is initiated by the removal of all those reactions that are exclusively associated with an enzyme of interest to determine the number of metabolites whose production the absence of the enzyme prevents. Damage analysis carried out by Lemke *et al.* on a graph representing *E. coli* metabolism showed that the extent of the damage relates to the importance of the enzyme [36]. They found that the loss of 91% of enzymes (one at a time) caused little damage to the network whilst 9% caused serious damage. Experimental results confirmed that this group contains the majority of enzymes that are essential to the viability of *E. coli*.

1.4.2 Null-space analysis

As described in Section 1.3.3, the null space of the stoichiometry matrix defines the metabolic capabilities of the system. It is therefore essential to define and analyse the kernel matrix to answer biological questions pertaining to the system under study. A number of analytical procedures exist, which are described in the rest of this section.

1.4.2.1 Dead reactions

Dead reactions [28] or strictly detailed balanced reactions [16] are those reactions in a metabolic model that can only have a zero rate at steady state. This applies whenever an orphan metabolite participates in that reaction. An example of a dead reaction is the reaction r_7 leading to the orphan metabolite F in Figure 1.7(a). Dead reactions can

[8] Also referred to as hubs, these are essential for the integrity of the network.

$$\begin{array}{c} t_1 \\ r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_6 \\ r_7 \\ t_2 \end{array} \begin{bmatrix} -2 & 0 \\ -2 & 0 \\ -1 & 1 \\ 0 & -1 \\ 0 & 1 \\ -1 & 0 \\ 0 & 0 \\ -1 & 0 \end{bmatrix} \leftarrow$$

Figure 1.10 – Matrix representing a zero row indicating the dead reaction r_7 in the null space of the stoichiometry matrix in Figure 1.7(b).

easily be identified from the kernel of the null space if their corresponding row is a zero row (Figure 1.10), indicating that they are incapable of carrying flux at steady state.

Identifying dead reactions in a reconstructed metabolic model is a vital step in detecting potential errors in the model definition, such as incorrect reaction stoichiometries and incomplete or inaccurate data (i.e. missing reactions due to incorrect annotation).

1.4.2.2 Enzyme subset analysis

Enzyme subsets (ES), or reaction subsets[9], are groups of enzymes in a metabolic model that operate together in fixed flux proportions at steady state [37]. Enzymes participating in an ES must fulfil the following two conditions: the ratio of any random pair of flux vectors satisfying the steady-state approximation has to be the same non-zero value, and the orientations of the irreversible reactions involved must not contradict each other. Typical examples include reactions in a linear system, where the flux through each of the reactions is equal in any steady-state flux distribution. An illustrative example of an ES in a slightly more complex system, represented by the simple metabolic model in Figure 1.2(a), is shown in Figure 1.11. If any reaction in this ES is removed from the model, then the other reactions in the subset cannot work properly and they will have a zero flux in steady state.

ESs can be identified from the null space of the stoichiometry matrix by finding the rows that only differ by a (scalar) factor. For example, the null space of the simple metabolic model in Figure 1.2(a) indicates the presence of six ESs, one containing three reactions and the other five containing one reaction each. Rows corresponding to the reactions r_1, t_1 and t_2, forming the larger ES, differ by a factor of two (Figure 1.11(b)). The other five ESs are formed by uncoupled rows in the kernel matrix. The algorithm for detecting enzyme subsets as outlined in [37] is given below:

[9] It is preferable to use the term reaction subset rather than enzyme subset as it is reactions, not enzymes that carry flux. Moreover, an enzyme may catalyse more than one reaction, all of which may not be part of the same subset.

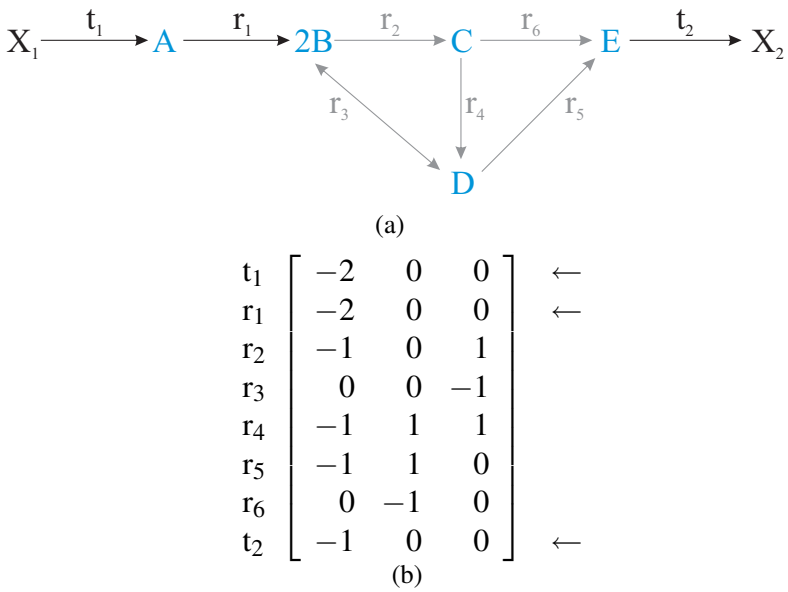


Figure 1.11 – Enzyme subsets (a) and its null-space matrix (b).

- Detect all row vectors of **K** that are null vectors (i.e. imply dead reactions that eventually end up in a single subset).
- Normalise each of the remaining row vectors of **K** by dividing by its greatest common divisor.
- Compare any normalised row vector with any other. If they are the same and there are no contradictions in the directionalities of irreversible reactions, the corresponding reactions belong to the same subset. The quotient of the normalisation factors gives the flux ratio.

ESs can be used to transform a complex metabolic model into a simplified model with equivalent topological properties by replacing the reactions involved in a subset by a single overall reaction [38]. This not only aids in the interrogation of the model but also simplifies it for more computationally intensive analysis methods (Section 1.4.3.1) [38]. Another important aspect is that enzymes belonging to the same subset can be assumed to share similar patters of genetic regulation. Schuster *et al.* used a model of the central metabolism of the yeast *S. cerevisiae* to study the correlation between enzyme subsets and microarray[10] expression data [38]. They showed that variation in the relative change of expression within the enzyme subsets is significantly lower when compared to the enzymes grouped randomly. Reed and Palsson [39] obtained similar results from a study performed on a model of *E. coli*. This property of enzyme subsets may suggest the regulatory structure of the metabolism [40, 11] and can aid genome annotation by suggesting missing enzymes from 'broken' subsets. ESs have been extensively used within our group to study their relationship with gene

[10] More on microarray data analysis can be found in Section 1.5.

expression [41]. It was observed that genes coding for reactions in a subset showed correlated changes in expression and that many subsets belong to known operons or regulons. In a later study ESs were used to simplify large metabolic models [42].

1.4.2.3 Reaction correlation coefficients and metabolic trees

The reaction correlation coefficient (RCC) describes the correlation between the fluxes carried by reactions at all possible steady states of the system [43]. It can be regarded as a quantitative extension of the qualitative concept of reaction subsets described in the previous section.

RCCs can be calculated from the null space **K** of the stoichiometry matrix. As discussed earlier in Section 1.3.3, the null space is spanned by the column of the $n \times d$ kernel matrix, where n is the number of reactions and d is the dimension of the null space. Each reaction is therefore associated with a d dimensional row vector (Equation 1.6). One apparent drawback of this approach is that **K** is non-unique and depends upon both the algorithm used for its calculation and the initial row and column order of **N**. However, the angles between the row vectors of any **K** for a given **N** are unique, provided **K** is orthogonal[11] [43]. Hence, RCCs are calculated from the orthogonal **K** matrix. Other modifications of **K** prior to the calculation of RCCs include the removal of zero row vectors (dead reactions), as no angle can be meaningfully assigned between a zero vector and any other. Moreover, isostoichiometric reactions are removed from the model, as they do not add any new information to the structural model and distort the results obtained from it [43].

The correlation between fluxes carried by reactions in the **K** matrix is measured by the cosine of the angle ($\cos(\theta_{ij}^{\mathbf{K}})$) between the row vectors $\mathbf{K}_i$ and $\mathbf{K}_j$, i.e.

$$\cos(\theta_{ij}^{\mathbf{K}}) = \frac{\mathbf{K}_i \mathbf{K}_j^T}{\sqrt{\mathbf{K}_i \mathbf{K}_i^T}\sqrt{\mathbf{K}_j \mathbf{K}_j^T}} \tag{1.8}$$

$\cos(\theta_{ij}^{\mathbf{K}})$, denoted with the symbol ϕ, is called the reaction correlation coefficient. Statistically it relates to the Pearson's (population) correlation coefficient[12], r_{ij}, between the fluxes carried by the pair of reactions i and j for all possible steady states of the system. Values of ϕ fall within the range $-1 \leq \phi \leq 1$. When ϕ is equal to ± 1, the rows $\mathbf{K}_i$ and $\mathbf{K}_j$ are parallel, which implies that they carry steady-state flux in a fixed ratio. This is equivalent to stating that these reactions are members of the same subset. Whereas

[11] That is, $\mathbf{K}\mathbf{K}^{\mathrm{T}} = \mathbf{I}$, in which case column vectors are orthogonal, and **K** represents an orthogonal basis of the null space of **N**.

[12] Pearson's correlation coefficient (r) [44] is a measure of the correlation (linear dependence) between two variables X and Y, giving a value between +1 and 1 inclusive. It is widely used in the sciences as a measure of the strength of linear dependence between two variables. Please see Section 1.5 for more on Pearson's correlation coefficients and correlation analysis.

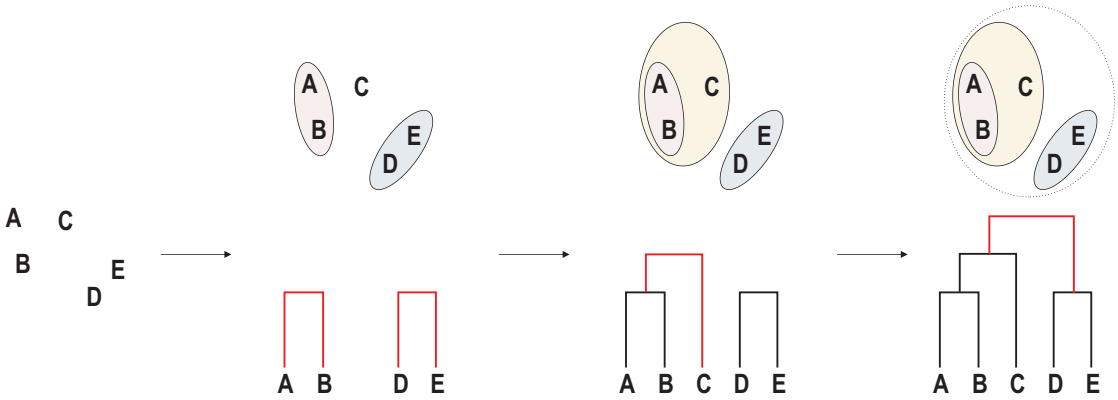


Figure 1.12 – Schematic representation of a hierarchical agglomerative clustering algorithm for five object A-E. See text for detailed description. Adapted from [45].

when ϕ is equal to 0, vectors $\mathbf{K}_i$ and $\mathbf{K}_j$ are orthogonal, stating that the reactions i and j are stoichiometrically disconnected subsystems [43].

Absolute values of RCCs can be used as a similarity measure for clustering reactions based on the correlation between fluxes carried by them. Clustering is one of the unsupervised approaches to classify data into groups with similar patterns that are characteristic to the group. There are many methods of clustering and discussing each one of them is beyond the scope of this thesis. Interested readers are referred to [45] for a detailed description of the various clustering methods. Nevertheless, it must be noted that throughout this thesis hierarchical agglomerative clustering method is used for classifying data. This method involves grouping objects into clusters and specifying relationships among objects in a cluster, resembling a phylogenetic tree. The first step in the procedure for hierarchical agglomerative clustering is the calculation of pairwise distance or similarity measures (in this thesis either Pearson's correlation coefficient or reaction correlation coefficient is used as the pairwise similarity measure) for the objects to be clustered. Based on the pairwise distance between them, objects that are similar to each other are grouped into clusters. Once this is done, pairwise distance between the clusters are recalculated, and clusters that are similar are grouped together in an iterative manner until all the objects are included into a single cluster [45]. See Figure 1.12 for a schematic representation of a hierarchical clustering algorithm. Several methods exist by which the distance between the clusters — or between clusters and objects — can be measured. In this study, however, the weighted (where the size of the cluster is accounted for) average distance between every point in a cluster and every point in the other cluster is taken as the distance between clusters. The clustering algorithm used here to achieve this is called the WPGMA (Weighted Pair Group Method using Arithmetic Averaging) algorithm [46]. The result of such clustering algorithms are dendrograms representing the various clusters in the data. Dendrograms are usually obtained in Newick format[13] and can be visualised using phylogenetic tree viewing programs such as NJPlot[†] [47] and

[13] Standard representation of a graph theoretical tree in text format using parentheses and commas.

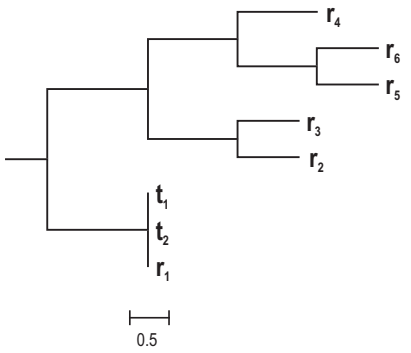


Figure 1.13 – Dendrogram representing the correlation between fluxes carried by reactions in the simple model described in Figure 1.2(a).

MEGA† [48] (software of choice). Other popular multipurpose phylogenetic software packages include freely-available PHYLIP† and SplitsTree† and commercial PAUP†.

A dendrogram generated by clustering RCCs of the simple metabolic model in Figure 1.2(a) using the WPGMA algorithm is shown in Figure 1.13. Note that the reactions that carry similar flux are clustered together. Poolman *et al.* used RCCs and metabolic trees to decompose large metabolic models into small functional modules [43]. In addition to their application in the construction of metabolic trees and modular decomposition, RCCs can be used as a convenient means of identifying disconnected subnetworks in metabolic models. RCCs are used later in this thesis to investigate a model of plant carbon metabolism and to integrate metabolic models with gene expression data.

1.4.3 Pathway analysis

Pathway analysis deals with the discovery and analysis of biologically meaningful routes (or pathways) in metabolic networks that define the capabilities of the system [14, 17, 49]. Reaction routes are represented by a set of basis vectors in the null space of the stoichiometry matrix. However, as discussed earlier, these vectors are not unique, as there may be many sets of vectors that can be used to span the null space that are both theoretically and biochemically feasible. For this reason, the reaction routes as determined by these vectors cannot be an invariant property of the network. Moreover, they do not take into consideration the irreversibility of some reactions in the network. Information about irreversibility defines the thermodynamic properties of the network.

To overcome this obstacle of the consideration of irreversibility constraints, the vector of fluxes, $\mathbf{v}$, is decomposed into two subvectors, $\mathbf{v}_{rev}$ and $\mathbf{v}_{irr}$, which include the fluxes of the reversible and irreversible reactions, respectively. Provided that the directions of irreversible reactions are appropriately defined, their flux rates are always positive:

$$\mathbf{v}_{irr} \geq 0 \tag{1.9}$$

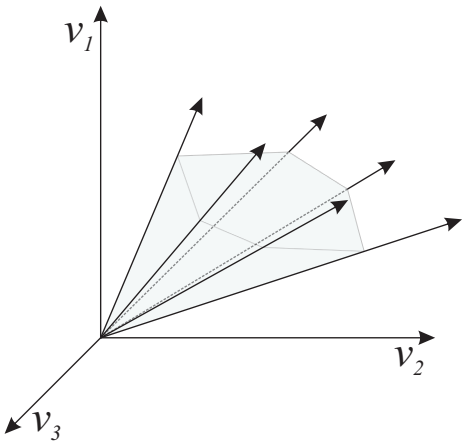


Figure 1.14 – Representation of the convex flux cone. The generating vectors represent fluxes and the polyhedron is to visualise the three-dimensional structure of the cone.

Any particular inequality, specified in the above relation (Equation 1.9) together with Equation 1.3, shrinks the region of admissible steady-state flux vectors to a subset of the null space in the positive orthant, referred to as the half-space of the null space of **N**. The intersection of all these half-spaces takes the shape of a convex polyhedral cone, called the *flux cone*, with a finite number of edges (Figure 1.14). Within this flux cone lie all the possible steady-state solutions and hence the flux distributions under which the system can operate. By virtue of the cone being convex, any vector within the cone can be represented as a non-negative linear combination of the generating vectors of the cone, which corresponds to its edges. Moreover, the edges of the cone are unique except for arbitrary scaling and correspond to biochemically feasible pathways [17].

To analyse the properties of flux cones for the discovery and analysis of pathways within metabolic networks, linear algebraic methods have been replaced with the mathematics behind the study of convex spaces[14] [17, 50]. Convex analysis is a branch of mathematics that can be used to analyse systems of linear inequalities - it can be used to overcome the shortcomings associated with reaction directionality and non-uniqueness [51]. A detailed description of two popular approaches that use algorithms developed using convex analysis to detect routes within metabolic networks forms the rest of this section.

1.4.3.1 Elementary modes analysis

The mathematics of convex analysis has been used in the development of algorithms for computing unique sets of generating vectors of convex polyhedral cones. These vectors are called *flux modes* (Figure 1.14). A flux mode can be described as a steady-state flux distribution in which the proportions of fluxes are fixed while their absolute magnitudes

[14] A convex space is one that satisfies the following condition: given any two points in the space, the line segment in between the points is completely contained in the space.

are indeterminate [50]. It is called an 'elementary (flux) mode' (EM) if, and only if, it satisfies the conditions [49, 50, 52]:

c1. Steady state: There is no net consumption or production of any internal metabolites, so that it satisfies the steady-state approximation (Equation 1.3).

c2. Feasibility: All fluxes in the mode are thermodynamically feasible, i.e. all irreversible reactions in the mode proceed in the appropriate direction (Equation 1.9).

c2. Non-decomposability: This means that the mode cannot be represented as a positive linear combination of other flux modes that satisfy the conditions c1 and c2 and are not zero vectors. In other words, a flux mode cannot be decomposed into further modes. Therefore, these modes represent *minimal* functional units within the network. Here, 'minimal' means that if only the reactions belonging to this set were operating, removal of one of these would lead to the cessation of any steady-state flux in the rest of the mode.

Thus, an elementary mode (EM) can be defined as a minimal set of reactions that could operate at steady state with all irreversible reactions proceeding in the appropriate direction [14, 49]. With this definition, the set of EMs in a metabolic network can have the following three properties [50, 53]:

i. There is a unique set of EMs for a given network. Therefore, this set is an invariant systemic property of the metabolic network being investigated.

ii. Each EM is genetically independent[15] or non-decomposable. That is, it consists of the minimum number of reactions that it needs to exist as a functional unit. If any reaction in an elementary mode were removed, the whole elementary mode could not operate as a functional unit.

iii. The elementary modes are the set of all routes through a metabolic network consistent with property ii.

Consider the metabolic network shown in Figure 1.2(a). The three EMs that occur in this network are shown in Figure 1.15 (EMs 1, 2 and 3). The pathway represented by EM 1 is built up by the reactions t_1, r_1, r_2, r_6 and t_2. It describes the uptake of external metabolite X_1 and the subsequent synthesis and excretion of X_2 via metabolites A, B, C and E. EM 1 fulfils the steady-state approximation as none of the internal metabolites has unbalanced production and consumption. It is thermodynamically feasible as all the irreversible reactions proceed in the appropriate direction. Finally, EM 1 is non-decomposable, as it cannot be decomposed into further modes. Similar properties characterise the other two EMs 2 and 3.

[15] Metabolic routes that are linearly dependent but defined as independent genotypes.

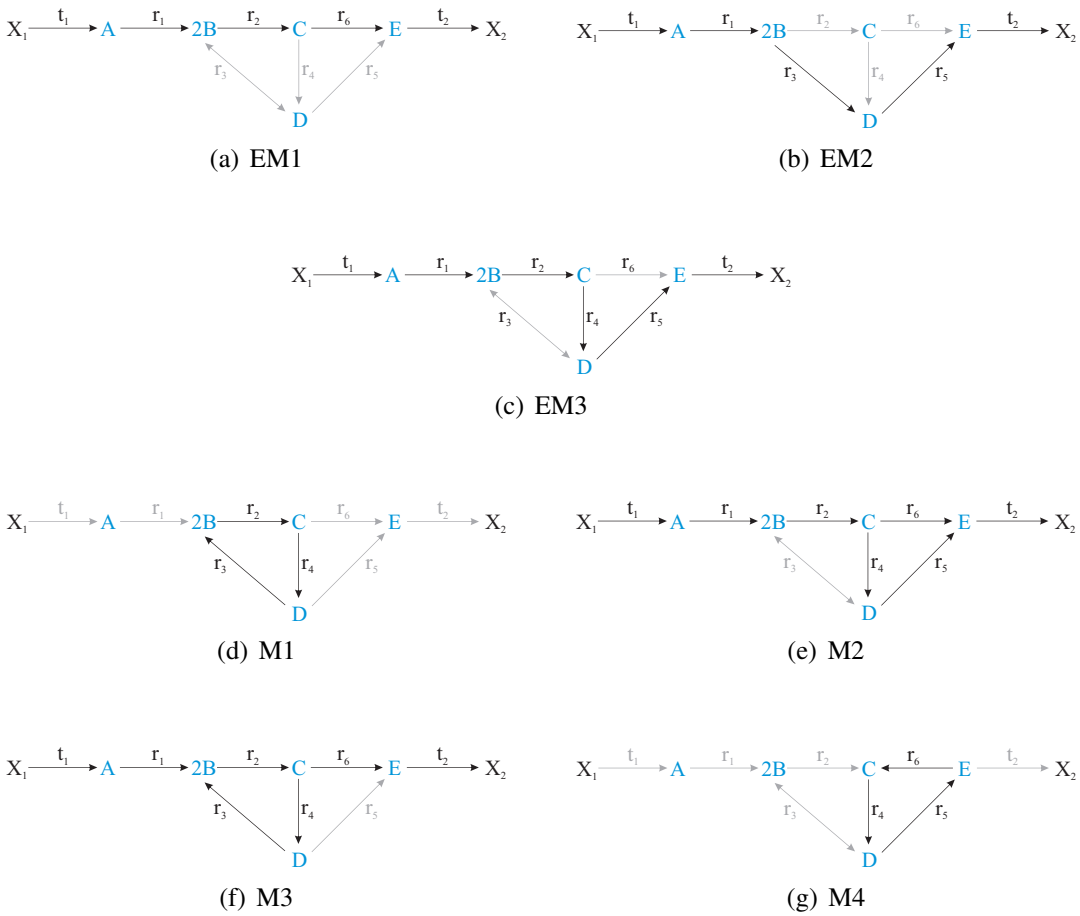


Figure 1.15 – Various modes in the metabolic network shown in Figure 1.2(a). EM1, EM2 and EM3 represent the three EMs, while routes M1 - M4 represent modes that are non-elementary.

On the other hand, consider the other possible routes through the metabolic network as shown by the remaining modes M 1 - M 4 (Figures (d)-(g)) in Figure 1.15. Here, mode M 1 is formed by complete balancing of only internal metabolites. Such modes are called *futile* (or substate) cycles. They are not generally considered as EMs (even though they satisfy conditions c1, c2 and c3) as they are thought to be biologically wasteful [14, 54]. However, if they are 'driven' i.e. there are other external metabolites being consumed then they are considered as genuine EMs. If no other substances are involved in these modes then there is no thermodynamic driving force and hence no net flux. Modes M 2 and M 3 are not elementary as they can be decomposed into further modes. For example, mode M 2 can be decomposed into EMs 1 and 3. Finally, mode M 4 is not an EM since it is a futile cycle and is not thermodynamically feasible (the irreversible reaction r_6 proceeding in an inappropriate direction).

Several algorithms have been proposed for the enumeration of EMs such as the canonical basis approach [49], the null-space approach [55] and the binary approach [56]. Despite these algorithmic advances, the computation of EMs for larger metabolic models meets the problem of combinatorial explosion in the requirement for

computational memory and processing power. The number of EMs was found to grow exponentially with increasing network size [57]. Because of this limitation, EM analysis has mainly been applied to networks of small or moderate size. However, the explosion of the number of EMs can be controlled by carefully managing the irreversible reactions, enzyme subsets and highly connected metabolites in the model. While the former reduce the number of EMs by imposing additional thermodynamic constraints to limit the size of the solution space, enzyme subsets help to reduce the load on the computer by making it possible to replace a set of reactions with a single reaction (see Section 1.4.2.2). Highly connected metabolites can be defined as external to split the network into sub-networks, which are easier and possibly more convenient to analyse [58].

EM analysis has proven useful in the interrogation of the properties of numerous metabolic models and has become an important theoretical tool for biotechnology and metabolic engineering to generate and test novel hypotheses [51, 56]. The number of EMs is an important measure that characterises the network's flexibility (redundancy[16], robustness) to perform a certain function. Meanwhile, the frequency of participation of a particular reaction in the set of EMs indicate the importance of that reaction for system performance under different growth regimes [59]. For example, if a reaction is involved in all the growth-related EMs its deletion can be predicted to be lethal, since all those EMs would disappear. A similar method is employed to infer the viability of mutants *in silico*.

An increasing number of metabolic networks have been studied, including that of the human red blood cell [60], *E. coli* [59], *H. influenzae* [61], *H. pylori* [11] and *S. cerevisiae* [62], to elucidate their topological properties and to identify novel pathways within them. Recently, EM analysis was applied to a model of Calvin cycle to study light/dark metabolism in plants [63]. A similar work analysed a model of plant mitochondrial TCA cycle to describe its structural properties [64]. In many of these studies, EM analysis was successfully employed to identify futile cycles operating within the network. It was also used to analyse the structural couplings between reactions, which might give hints for underlying regulatory circuits, such as the enzyme/reaction subsets [58].

In addition to the application of EM analysis to elucidate the properties of biochemical networks, it has been used in the field of metabolic engineering for more practical purposes. EM analysis was used to analyse a model of the central metabolic reactions of *E. coli* to predict optimal and sub-optimal yields of aromatic amino acids from carbohydrate substrates. A strain of *E. coli* was then engineered which achieved the predicted yield values [65]. Recently, EM analysis applied to a model of *S. cerevisiae* intermediary metabolism was then used in a recombinant strain to study the effect of gene additions and deletions on the yield values of the production

[16] The number of independent pathways in the system that have equivalent input and output fluxes.

of poly-β-hydroxybutyrate [62]. A similar work on *S. cerevisiae* describes the use of EM analysis to assign function to orphan genes [12]. For more on the applications of EM analysis, interested readers are directed to [51] and [56].

1.4.3.2 Extreme pathway analysis

While EM analysis enumerate all distinct routes within a metabolic network, extreme pathway analysis focuses on enumerating the unique and minimal set of convex basis vectors needed to describe all the possible steady-state flux distributions in the network [66]. Thus, extreme pathways (EPs) define the edges of the convex cone (generating vectors) that represent all the possible flux distributions in the metabolic network and are a subset of EMs [51]. Although EPs share properties 1 and 2 in Section 1.4.3.1 with EMs, they differ from EMs as they have to satisfy an additional property of systemic independence [66]. A set of EPs {p1, ..., p2} is said to be systemically independent if no EP can be written as a non-trivial non-negative linear combination of any other EP. Note that the difference between this definition and linear independence is that the coefficient of linear combination cannot be negative [66]. Therefore, the set of EPs does not include all genetically independent routes, as it is the case with EMs, instead contains a subset still capable of spanning the convex cone.

The algorithms for the calculation of EPs and EMs differ in terms of the treatment of reversible and irreversible reactions in the model. EP analysis decouples all internal reversible reactions into two separate reactions for the forward and reverse directions, and subsequently calculates the pathways. Since EPs are a subset of EMs, considering EPs instead of EMs reduces not only the number of routes (particularly important in case of large metabolic networks) but also the computational power required [67, 53]. However, the use of EP analysis to study system properties demands careful consideration as it does not produce the complete set of genetically independent routes within the metabolic network under consideration. For the same reason, structural robustness and the relative importance of reactions cannot be assessed properly [67].

Nevertheless, EP analysis has been used in the analysis of a broad range of metabolic models. It was used to demonstrate that the EP structure in a model of *H. influenzae* shows significant network redundancy when compared to that of *H. pylori* [68]. A similar work on *H. pylori* studied amino acid production to analyse the emergent properties of the system [69]. Recently, Wiback and Palsson (2002) applied EP analysis to a model of human red blood cell metabolism to study its physiology and to demonstrate that EPs can be used to interpret the steady-state solution space with respect to network capabilities [70].

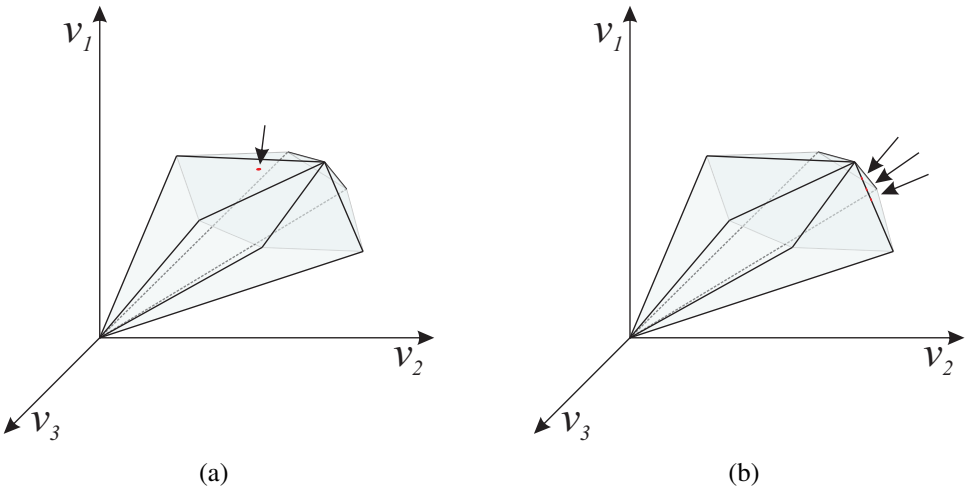


Figure 1.16 – Flux analysis (a) In metabolic flux analysis measured fluxes are combined with the stoichiometric constraints and are used to shrink the solution space. A possible ideal solution representing the actual flux distribution is indicated by the arrow. (b) Physiochemical constrains along with possible maximum and minimum fluxes through any particular reaction is used to constrain the solution space in flux balance analysis. Multiple optimal solutions in some systems are indicated by red dots on the edge.

1.4.4 Metabolic flux analysis

In several methodologies within the stoichiometric modelling framework, steady-state metabolic models are coupled with available experimentally measured fluxes to completely determine the current flux distributions in the system (Figure 1.16(a)). This approach is called metabolic flux analysis (MFA). Here, constraints imposed by a set of measured fluxes are used to shrink the possible solution space defined by Equation 1.3 to determine those fluxes that have not been measured (Figure 1.16(a)) [71]. This is attained by partitioning the steady-state rate equation to accommodate the measured (index m) and unknown fluxes (u):

$$0 = \mathbf{N}\mathbf{v} = \mathbf{N}_u\mathbf{v}_u + \mathbf{N}_m\mathbf{v}_m \tag{1.10}$$

$$\mathbf{N}_u\mathbf{v}_u = -\mathbf{N}_m\mathbf{v}_m \tag{1.11}$$

An ideal solution to this equation is a unique point in the null space of $\mathbf{N}$ representing the actual flux distribution. This happens only when $\mathbf{N}_u$ is a square matrix (number of unknown reactions is equal to the number of unknown metabolites) and invertible because then all unknown rates in $\mathbf{v}_u$ can be determined. On the contrary, the system is underdetermined when at least one unmeasured flux, and probably most of the fluxes, are not calculable [72, 25], which is often the case in reality. An algorithm for the calculation of all the possible fluxes in an underdetermined system was developed by Klamt *et al.*, the application of which was illustrated using a model of the central metabolism in purple nonsulphur bacteria [72]. Interested readers are directed to [71] and [72] for more extensive description of the basic techniques of MFA.

MFA along with C^{13} labelling has been used widely to characterise canonical and physiological states of cells in batch and continuous cultures [73]. Animal [74] and plant cell cultures [75] have been extensively studied using MFA techniques. It has also been applied to study transient processes in microorganisms [76].

1.4.5 Flux balance analysis

Flux balance analysis (FBA) [77, 10] seeks to identify physiologically meaningful and optimal flux distributions in underdetermined metabolic networks. Here, the solution space defined by the mass balance constraints (Equations 1.1 and 1.3) is further restricted by imposing the invariant and adjustable physiochemical constrains given in Table 1.1 [78] and by specifying the maximum and minimum fluxes through any particular reaction (Figure 1.16(b)) [79]. The addition of these constraints results in the definition of a bounded solution space wherein every possible flux distribution, or every possible metabolic phenotype of the cell must lie. In order to determine the actual flux distribution within the resulting constraint-defined space of feasible distributions, it is assumed that cells have evolved to achieve an optimal behaviour owing to evolutionary pressure. This allows the underdetermined system to be formulated as an optimisation problem[17]. If the objective function[18] is linear[19], the optimisation problem is a linear programming (LP) problem [79].

Solution to the LP problem is a single flux distribution through the bounded flux cone that can be used to investigate the metabolic capabilities of the system. However, this flux distribution is highly dependent on the constraints specified in the objective function and may not correspond to the actual flux distribution. By calculating and examining optimal flux distributions under various conditions, it is possible to generate quantitative hypotheses *in silico* that may be tested experimentally.

Predictions by FBA have been shown to be consistent with experimental data in 86% of instances for *E. coli* [80] and in 60% instances for *H. pylori* [11]. FBA was applied to a model of *H. influenzae* to show that alternate optimal solutions can be used to find redundancies in the metabolic network [68]. A number of FBA studies involve *E. coli* as the organism of choice as it is one of the few organisms for which there is a genome-scale model and a large body of experimental evidence. Once such study employed a quadratic objective function to improve the prediction efficiency of FBA [81]. This approach is called minimisation of metabolic adjustment (MOMA) and it aims to find a point in the solution space that is closest to an optimal point in the wild-type solution space. MOMA was used to demonstrate that genetically engineered knockout undergo a

[17] For example, the maximisation of biomass production or the minimisation of ATP utilisation.

[18] A function to be maximised or minimised in optimisation theory.

[19] A function f of variables x_1,...,x_n is called a linear form if it can be written as $c_1x_1 + ... + c_nx_n$, where the coefficients c_i are constant real numbers.

minimal re-distribution with respect to the flux configuration of the wild-type cell [81]. Recently, attempts were made to use FBA on whole-plant models. Unfortunately, it was found that the characteristics of plant metabolism such as the compartmentation, make this task very difficult [82].

1.5 Integration of gene expression data into stoichiometric models

In the previous sections describing the basic principles and techniques involved in the construction and analysis of stoichiometric models, it was shown that the information contained in a stoichiometric model itself results in an underdetermined linear equation system (Section 1.3.3). This information is not sufficient either to calculate a unique flux distribution or to understand the genetic and metabolic regulations within the system. Such models can, therefore, be combined with additional experimental data to provide a deeper insight. One set of data that can be readily incorporated with stoichiometric models is the experimentally measured flux values. Two of the most popular techniques that involve combining flux values with stoichiometric models to completely determine the current flux distributions in a system were described in Sections 1.4.4 and 1.4.5.

Another set of experimental data that may be effectively integrated with stoichiometric models is the expression levels of genes coding for reactions in the model. This information can be used to understand the important features of genetic and metabolic regulation with the system. Expression levels of all genes in an organism can be measured using a high-throughput functional genomic technology called DNA microarrays. Microarrays may be used to measure gene expression in many ways, a complete description of which is outside the scope of this thesis. However, the general procedures involved in a typical microarray experiment are shown in Figure 1.17. One of the most popular applications of microarrays is to monitor the expression level of genes at a genome-scale. Patterns could be derived from analysing the change in expression of the genes, and new insights could be gained into the underlying biology [84]. The first step here is to process the data obtained from the image generated at the end of the microarray experiment (Figure 1.17). Numerous statistical methods are available to do this, some of which are described in [83]. The final processed data can be represented in the form of a matrix, referred to as the gene expression matrix (Table 1.2). Each row in the matrix corresponds to a particular gene and each column corresponds to an experimental condition. Expression levels of a gene across different experimental conditions are called the gene expression profile, and the expression levels of all genes under an experimental condition are called the sample expression profile [84]. An example of such a database that stores the expression profiles of all genes in an

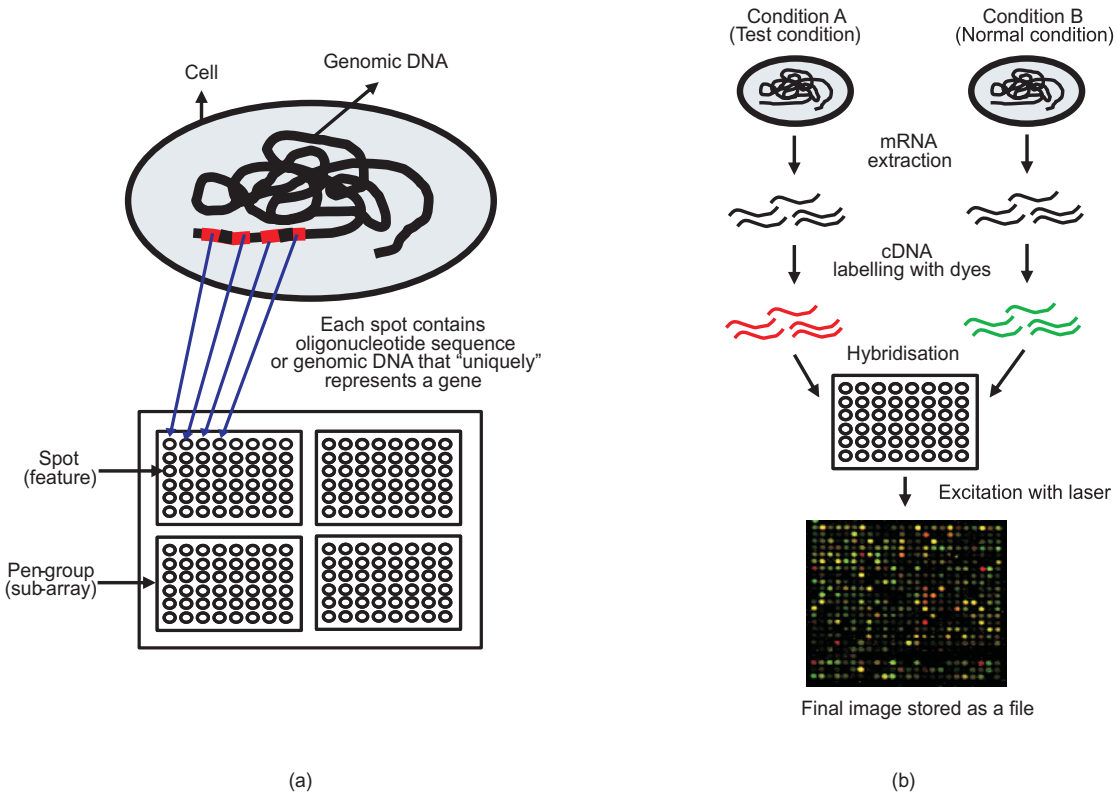


Figure 1.17 – (a) A microarray may contain thousands of 'spots. Each spot contains many copies of the same DNA sequence that uniquely represents a gene from an organism. Spots are arranged in an orderly fashion into Pen-groups. (b) Schematic of the experimental protocol to study differential expression of genes using the two dye comparative methodology. The organism is grown in two different conditions (a reference condition and a test condition). RNA is extracted from the two cells, and is labelled with different dyes (red and green) during the synthesis of cDNA by reverse transcriptase. Following this step, cDNA is hybridised onto the microarray slide, where each cDNA molecule representing a gene will bind to the spot containing its complementary DNA sequence. The microarray slide is then excited with a laser at suitable wavelengths to detect the red and green dyes. The final image is stored as a file. Interested readers are referred to [83] for a complete description of the methods involved in the extraction of data from these image files. Reproduced from [84].

organism under different experimental conditions is the Nottingham *Arabidopsis* Stock Centre's (NASC) microarray database - NASCArrays[†] [85]. It contains more than 3000 hybridisations, each with expression level measurements for over 22,500 genes represented on the *A. thaliana* ATH1 arrays[20]. These arrays are derived from varied experiments, tissues, conditions, treatments and genetic backgrounds.

Analysis of data in the gene expression matrix is based on the comparison of gene expression or sample expression profiles. A variety of distance measures are used to calculate the similarity in expression profiles, one of the most commonly used among which is Pearson's correlation coefficient (denoted by the symbol r). Values of r range from +1 to -1. In general, the correlation expresses the degree that, on an average, two

[20] The GeneChip® Arabidopsis ATH1 Genome Array contains more than 22,500 probes synthesised *in situ* and designed to measure temporal and spatial gene expression in approximately 24,000 gene sequences

Table 1.2 – A gene expression matrix.

	Condition A	Condition B	Condition C	Condition D	Condition ...
Gene 1	2521.14	264.23	456.5	456.34	...
Gene 2	156.5	2164.45	2456.56	567.56	...
Gene 4	1567.12	2634.63	234.51	12.55	...
Gene 15	1483.64	324.67	678.65	423.77	...
Gene 61	56.15	264.89	23.34	556.7	...
Gene 34	8.5	64.21	56.67	43.13	...
Gene ...	...	...	...	...	...

variables change in concert. An r value closer to +1 indicates a positive correlation where the expression level of a gene increases when the expression level of another gene increases. On the contrary, a value closer to -1 indicates a negative correlation where the genes have opposite expression profiles. Meanwhile, 0 means that no relationship can be inferred between the the expression profiles of genes. An example of a correlation matrix is shown in Table 1.3. The confidence in the correlation is quantified by the p value; it is the probability that the observed value of r could have been obtained by chance under the null hypothesis that the two variables being compared vary independently. If this p value is lower than the conventional 5% chance of the null hypothesis being true (i.e. $p<0.05$) then the correlation coefficient is considered to be statistically significant, which means that the null hypothesis is wrong and that there exists a real correlation between the expression profiles.

In order to combine metabolic models with gene expression profiles, however, it is necessary to identify the genes coding for the entire set of reactions in the model. This is a very difficult task as intricate associations exist between genes, proteins, enzymes and reactions. Each enzyme has a relatively high specificity in terms of the metabolites on which it can act. Most enzymes usually catalyse only a single reaction so that the products of the reaction are strictly determined. As with all proteins, enzymes are encoded by genes that make up an organism's DNA. In the trivial case, one gene encodes one enzyme that catalyses a single reaction. In most cases, however, the relationships between genes, proteins, enzymes and reactions are much more complex. Many enzymes can accept several different substrates thus relating one or more genes coding for this enzyme to several reactions. For reactions catalysed by enzyme complexes, the opposite situation applies where several genes are related to one reaction. Similarly, there are enzymes that differ in amino acid sequence but catalyse the same chemical reaction. Such enzymes, called isozymes, often have different kinetic parameters and regulatory properties associated with them. Under such a scenario, mapping the associations between genes, proteins, enzymes and reactions is often achieved with the help of public genome/pathway databases. An example of such a database that contains information about both predicted and experimentally determined pathways, reactions, compounds, genes and enzymes and the associations between them is the BioCyc

Table 1.3 – Correlation matrix representing the correlation between the expression profiles of genes shown in Table 1.2.

	Gene 1	Gene 2	Gene 4	Gene 15	Gene 61	Gene 34	Gene ...
Gene 1	1.0	-0.70	0.16	0.97	-0.45	-0.95	...
Gene 2	-0.70	1.0	0.13	-0.5	-0.29	0.86	...
Gene 4	0.16	0.13	1.0	0.055	-0.22	0.04	...
Gene 15	0.97	-0.58	0.05	1.0	-0.60	-0.90	...
Gene 61	-0.45	-0.29	-0.22	-0.60	1.0	0.21	...
Gene 34	-0.95	0.86	0.049	-0.90	0.21	1.0	...
Gene ...	...	...	...	...	...	...	1.0

database†. It is a collection of 505 pathway/genome databases. Each database in the BioCyc collection describes the genome and metabolic pathways and their associated enzymes of a single organism. An independent genome/pathway database that has similar structure and organisation as the BioCyc database is the AraCyc† [86] database for the model plant *A. thaliana*.

A number of studies have attempted to integrate the properties of microarray gene expression profiles with those of the steady state stoichiometric models. Correlation between ESs and gene expression profiles were investigated by Schuster *et al.* [38] using a steady state model of the central metabolism of *S. cerevisiae*. They showed that variation in the relative change of gene expression within an ES is lower when compared to enzymes that were grouped randomly [38]. From these results it was evident that ESs in metabolic models can provide insight into regulatory strategies as the genes that encode the corresponding enzymes in an ES are likely candidates for co-regulation. Reed and Palsson [39] obtained similar results from a study performed on a genome-scale model of *E. coli*. A recent study demonstrated correlated changes in the expression of genes coding for enzymes in an ES [41].

1.6 Software for analysing structural models

Construction and analysis of metabolic models is a tedious and error-prone task that needs to be repeated whenever the model is altered (Section 1.1). This process becomes increasingly complex and time-consuming with increase in the size of the model to be investigated. The use of high-throughput techniques and the holistic approach in systems biology means that most biochemical models may typically contain in excess of 50 reactions. Therefore, with the exception of the most trivial of cases, the use of software tools in metabolic modelling is highly indispensable.

Over the past decade, a number of software packages have been developed for the construction and analysis of metabolic models. One apparent criterion based on which these tools can be classified is their mode of interaction with the modeller [87]. While tools like COPASI† [88] (formerly GEPASI† [89]), YANA† [90] and FluxAnalyzer† [91]

employ a graphical user interface (GUI); METATOOL† [37], JARNAC† [92] (a post genomic version of SCAMP† [93]), PySCeS† [94] and ScrumPy† [95, 63, 96] use a command line interface (CLI). Though GUI based packages are easier for the novice user to learn and get accustomed to, CLI applications provide more user interaction, flexibility and extensibility[21]. Nevertheless, the distinction between these divisions are breaking down as most of these tools can now interact through a common model definition language - the Systems Biology Markup Language† (SBML) [97].

Many of the modelling tools discussed earlier, including COPASI and JARNAC, are designed for simulating metabolic networks on the basis of kinetic descriptions. Analysis of the underlying stoichiometry of a metabolic network has been considered only to a minor extent. METATOOL was one of the very first tools dedicated solely to stoichiometric analysis. It is a CLI based, reliable and high-performance implementation of the EM analysis algorithm coded in the C† programming language. METATOOL was later succeeded by a GUI based Java† application called YANA† [90], that has additional capabilities such as predicting a valid EM activity pattern from a given flux distribution.

Other popular structural modelling tools include CellNetAnalyzer† [98] (successor of FluxAnalyzer), PySCeS and ScrumPy. The former incorporates metabolic modelling capabilities to the commercially available numerical computing environment MATLAB† by facilitating structural and functional analysis of metabolic networks. PySCeS and ScrumPy are CLI based applications written in the Python† [99, 100] programming language (Section 1.6.1.1). Though both these tools have the added advantage of the capabilities inherited from Python, PySCeS provides little support for structural modelling (EMs are calculated by way of an interface to METATOOL) [94]. ScrumPy, on the other hand, has equal support for both kinetic and structural modelling. Built in our lab, it has been upgraded over the years with numerous functionalities for structural modelling. ScrumPy is open source[22] and is currently available for various Linux platforms. A brief description of some of the properties of ScrumPy along with intructions for basic usage follow.

1.6.1 Metabolic modelling with Python and ScrumPy

1.6.1.1 The Python programming language

Python is a remarkably powerful dynamic programming language that is used in a wide variety of applications. Though Python is a relatively modern programming language, it is often compared to C†, C++† or Java†. It is an open source and platform

[21] Ability to add new capabilities to existing software without major changes in its implementation.
[22] An open-source license makes a software freely usable and distributable, even for commercial use.

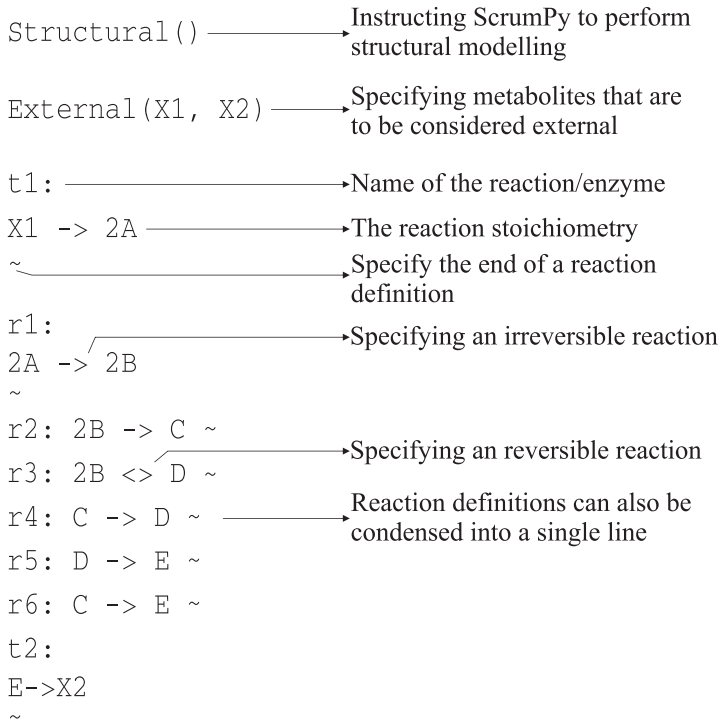


Figure 1.18 – Syntax of the ScrumPy '.spy' input file representing the simple metabolic model in Figure 1.2(a).

independent[23] language that is easy to learn, due to its very clear, readable syntax and console based interactive development environment. Python comes with an extensive library of statistical, numerical and scientific tools, such as SciPy† and NumPy†, that contain mathematical algorithms required for scientific and engineering applications. Python scripts can easily communicate with other parts of an application with a variety of integration mechanisms. Such integrations allow Python to be used as a product customisation and extension tool. Python code can invoke C and C++ libraries, can be called from C and C++ programs and can integrate with Java components [99, 100].

Python has deep support for software reuse mechanisms such as object-oriented programming (OOP). It is a programming paradigm that uses 'objects' to design computer programs. An object may contain data and/or instructions that operate on the data. In OOP, an object is an instance of a 'class'. The class object contains a combination of data and the instructions that operate on these data, making the object capable of receiving messages, processing data, and sending messages to other objects. Object functionality is defined by creating 'methods' within the class structure. Once the class has been instantiated (e.g. `instance()`) methods can be called using the 'dot' notation (e.g. `instance.method(argument)`; where `argument` is a variable or value passed into the method). With Python it is easy to write complex object-oriented programs that can be reused.

[23] Runs on all major operating systems: Windows, Linux/Unix, OS/2, Mac, among others.

Type	Description	Syntax Example
`str`	An immutable[a] sequence of characters	`'This is a string'`
`int`	Integer	`12`
`float`	Floating point	`3.141592`
`bool`	Boolean	`True` or `False` (`1` or `0`)
`tuple`	Immutable, can contain mixed types	`('string', 2.3, True)`
`list`	Mutable, can contain mixed types	`['string', 4.5, True]`
`dict`	Group of key and value pairs	`{'key1': 'a', 'key2':['no', 3.2]}`

[a] an object whose state cannot be modified after it is created.

Table 1.4 – Python built-in data structures with examples [99].

Python boosts developer productivity many times beyond compiled or statically-typed languages such as C, C++ and Java. Python code is typically 1/3 to 1/5 the size of equivalent C++ or Java code and it runs immediately without the lengthy compile and link steps required in these tools [99, 101]. Unlike the other programming languages, Python automatically allocates and reclaims ('garbage collects') objects when not in use. This property relieves the programmer from keeping track of the low-level memory details of data structures[24]. Data structures provided by Python as an intrinsic part of the language are string, integer, float, boolean, lists, tuples and dictionaries (Table 1.4). They are both flexible and easy to use.

For a more detailed description of the features and capabilities of Python as a programming language and to actually learn programming in Python, readers are referred to any recent edition of [99].

1.6.1.2 The ScrumPy metabolic modelling tool

It is possible to build and analyse metabolic models directly using only Python and SciPy [102]. Although flexible, this approach does require considerable skill in both numerical analysis and computer programming [94]. ScrumPy has been developed to provide a high-level modelling interface that utilises and extends the low-level capabilities provided by Python and SciPy, making it unnecessary for the modeller to work with advanced programming techniques or low-level numerical algorithms.

The ScrumPy package and detailed instructions for installing it in popular linux based operating systems is available for download from the main ScrumPy website (`http://mudshark.brookes.ac.uk/index.php/Software/ScrumPy`). Although a complete documentation detailing the usage of ScrumPy is available as part of the distribution, a very basic demonstration of its usage in constructing and analysing structural models is provided below.

[24] A data structure is a particular way of storing and organising data in a computer so that it can be used efficiently.

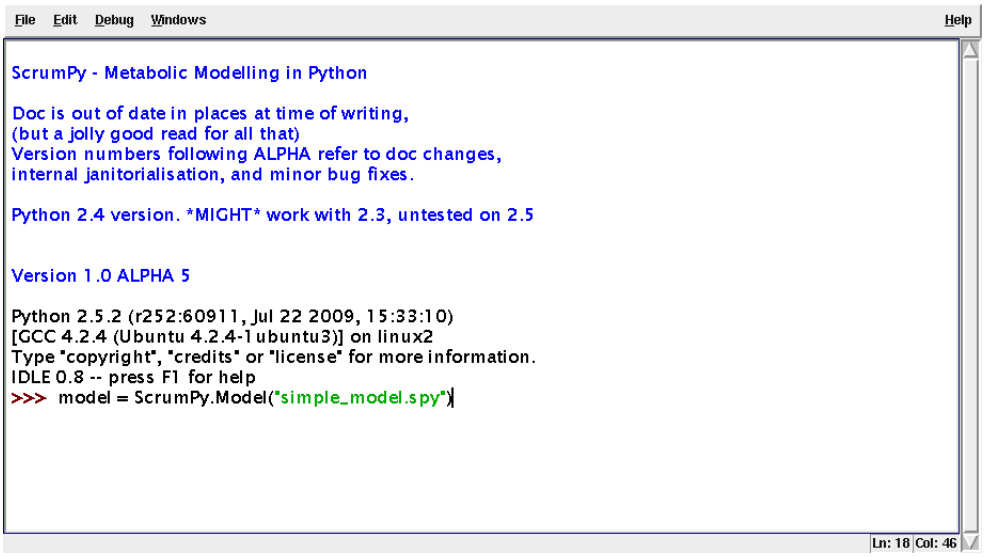


Figure 1.19 – The ScrumPy CLI.

Using ScrumPy

The first step in the modelling process is to create a model object from an input file. As a common practise within our lab, ScrumPy input files are denoted with a '.spy' file extention (e.g. `simple_model.spy`). Further information regarding the syntax of the input file, which describes a model in terms of its stoichiometry, can be found in Figure 1.18. The input file is loaded into ScrumPy by typing into the CLI (Figure 1.19):

```
>>> model = ScrumPy.Model("simple_model.spy")
```

The stoichiometry matrix and null space of the model can be displayed by using the commands:

```
>>> model.sm
```

and

```
>>> model.sm.NullSpace()
```

respectively. Errors in the input file are displayed in an error message window and highlighted in the model editor window. If the model description is changed in any way, it can be recompiled by:

```
>>> model.Reload()
```

Once the model object is instantiated, its structural properties (such as the stoichiometry matrix and the kernel matrix) are available as model attributes that can be used in further calculations. Such attributes of the model can be obtained by typing in:

$$
\begin{array}{c}
\\
\text{ElMo\_0} \\
\text{ElMo\_1} \\
\text{ElMo\_2} \\
\text{ElMo\_3}
\end{array}
\begin{array}{c}
\begin{array}{cccccccc} \text{r}_1 & \text{r}_2 & \text{r}_3 & \text{r}_4 & \text{r}_5 & \text{r}_6 & \text{t}_1 & \text{t}_2 \end{array} \\
\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{array}
$$

(a)

$$
\begin{array}{c}
\\
\text{ElMo\_0} \\
\text{ElMo\_1} \\
\text{ElMo\_2} \\
\text{ElMo\_3}
\end{array}
\begin{array}{c}
\begin{array}{cc} \text{X1} & \text{X2} \end{array} \\
\begin{bmatrix} -1 & 1 \\ -1 & 1 \\ -1 & 1 \\ 0 & 0 \end{bmatrix}
\end{array}
$$

(b)

Figure 1.20 – (a) Elementary modes reaction matrix (E_M) and (b) Elementary modes stoichiometry matrix (E_S). Elements of the matrices indicate fluxes and their direction through reactions in EMs.

```
>>> dir(model)
```

These attributes include most of the common model analysis tasks. For example, ESs in the model can be obtained as a python dictionary by typing:

```
>>> enzyme_subsets = model.EnzSubsets()
```

The results from most structural analyses using ScrumPy are presented in the form of matrices. For instance, a matrix representing all EMs in the network can be generated from the stoichiometry matrix using the command:

```
>>> elementary_modes = model.ElModes()
```

From the resulting EMs matrix, elementary modes reaction matrix E_M (EM in rows and reactions in columns) and elementary modes stoichiometry matrix E_S (elementary modes in rows and external metabolites in columns) can be obtained. While E_M represent the reactions in an EM and its associated flux (Figure 1.20(a)), E_S indicate the net usage of external metabolites by a given EM (Figure 1.20(b)). E_M and E_S can be obtained using the following commands:

```
>>> em = elementary_modes.mo
>>> es = elementary_modes.sto
```

As has been demonstrated here, the interactive and user-friendly nature of ScrumPy aims to integrate the numerous capabilities of Python programming language into the construction and analysis of metabolic models. Further information on metabolic modelling using ScrumPy can be found in the documentation. However, in the future chapters of this thesis, the reader will be provided with more updates on ScrumPy commands wherever deemed necessary.

CHAPTER 2

Introduction to modelling plant metabolism

2.1 Introduction

The previous chapter described the basic concepts and formalisms involved in the stoichiometric modelling of biochemical systems and how the various techniques involved in analysing the stoichiometry of metabolic models can be used to explain and investigate the biological properties and capabilities of the system. It is evident that the foremost task in any modelling effort is to accurately define the model under consideration with the help of the available genomic, molecular and metabolic information. For this reason, most of the model analysis endeavours have concentrated on organisms such as *E. coli* and *S. cerevisiae*, for which sufficient information is available to accurately describe the biochemical networks. However, the recent advent of high-throughput techniques has ensured that the genomic, molecular and metabolic information available for more complex eukaryotes such as plants, especially the model plant *Arabidopsis thaliana*, is beginning to rival that of *E. coli* and *S. cerevisiae* [103].

Apart from being the primary source of food, either directly or indirectly, plants are extensively being used as a means of producing sustainable raw materials such as fat and starch for industrial purposes. They are also the basis of production of numerous pharmaceuticals and biodegradable substances. The recent surge in the availability of data pertaining to plant metabolism has encouraged a number of kinetic modelling efforts aimed at increasing the productivity and/or feasibility of such industrial and pharmaceutical processes [104]. A few of these efforts will be briefly described elsewhere in this chapter.

Although application of stoichiometric modelling to plant systems is still in its infancy, a number of telling efforts have already been made. One major objective of this thesis is to construct and analyse such a model of plant central carbon metabolism to further understand its properties and behaviours. This chapter will provide an overview of the biochemistry of plant metabolism in view of the aspects investigated in this study. The chapter will conclude with a brief review of the various attempts to model plant metabolic networks (primarily stoichiometric) and their outcome.

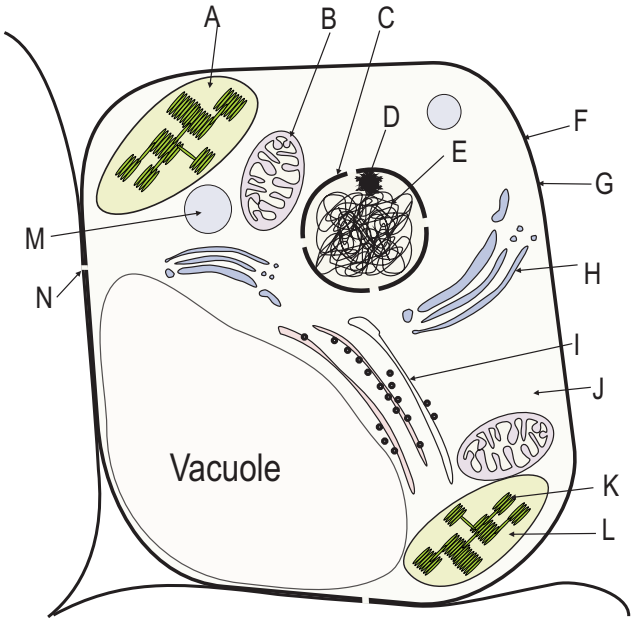


Figure 2.1 – Structure of a plant cell. A. Chloroplast, B. Mitochondria, C. Nucleus, D. Nucleolus, E. DNA, F. Plasma membrane, G. Cell wall, H. Golgi vesicles, I. Endoplasmic reticulum, J. Cytosol, K. Thylakoids, L. Stroma, M. Peroxisomes, N. Plasmodesmata

2.2 The biochemistry of plant metabolism

Unlike the single-celled prokaryotes, eukaryotic cells are structurally more complex and are characterised by a membrane-bound *nucleus* and other complex intracellular structures. Plant cells are eukaryotic cells that in particular have a very complex structure and differ in several key respects even from the cells of other eukaryotic organisms. The overall structure of a plant cell may be divided into *cell wall* and *protoplast*. The cell wall is composed of cellulose microfibrils embedded in a matrix of hemicellulose and pectin, and in many cases also lignin. It imparts plant cells with the necessary shape and structure required to provide the whole plant with mechanical support and protection from the external environment. The protoplast, on the other hand, contains the functional components of the plant cell. It is enclosed by a lipid bilayer impregnated with globular proteins called the *plasma membrane*. The contents of a protoplast can be divided into nucleus[1] and *cytoplasm* [105]. The latter contains a 'solution space' or matrix called the *cytosol* that supports other particles (e.g. ribosomes, mitochondria) and membrane systems[1] (e.g. endoplasmic reticulum (ER), golgi apparatus) in the cell (Figure 2.1) [105, 106].

The particles in the cytosol are either ribosomes[1] (generally considered as separate entities within the cell) or membrane-bound organelles commonly referred to as *compartments*. Compartments can be defined as reaction spaces enclosed by membranes

[1] The scope of this thesis restricts any detailed review of these systems. Interested readers are directed to [105] and [106].

Table 2.1 – Major metabolic compartments and their function in a plant cell.

Compartments/Subcompartments	Main Function
Cytosol	Supports other compartments
Plastids	Photosynthesis and storage of starch
Amyloplast	Storage of starch
Chloroplast	Photosynthesis
Stroma	Site for Calvin cycle
Thylakoid membrane	Site for light reactions
Lumen	Site for water splitting
Endoplasmic reticulum	Modification of new proteins and lipids
Golgi apparatus	Sorting and modification of proteins
Mitochondrion	Energy production
Matrix	Site for TCA cycle, β-oxidation
Vacuole	Storage of waste products and toxic materials
Nucleus	DNA maintenance and transcription

that perform specific functions within the cell. They act by sequestering the enzymes and metabolites participating in specific metabolic processes and thereby preventing the simultaneous occurrences of potentially incompatible reactions elsewhere within the cell [107, 108]. A list of various compartments and their respective functions within the cell can be found in Table 2.1. An illustration is given in Figure 2.1.

Based on the comparison of their contents, compartments are divided into two. They are *plasmatic* compartments that contain a high proportion of proteins (enzymes) and the protein-poor *non-plasmatic* compartments (e.g. vacuoles). Examples of plasmatic compartments include cytosol, plastid and mitochondria. Vacuoles, ER and golgi are, however, non-plasmatic. Many enzymes and metabolites in the plant cell are restricted to specific plasmatic compartments. This specificity is achieved either by the inability of certain molecules to penetrate the compartment membrane (e.g. ATP/ADP cannot penetrate the chloroplast membrane) or by being bound to structures within the compartments (e.g. ferredoxin bound to the thylakoid membrane) [106].

However, a number of examples exist in which the product of reactions located in one compartment is then utilised in another. Here, physically separating enzymes concerned with production of a substance from those involved in consumption allows both the processes to be regulated separately. This division of labour means that the metabolism in every compartment depends on other compartments for supplies of energy and metabolic precursors. Therefore, to understand the metabolism in plant cells it is necessary to know not only how metabolism and other processes are compartmentalised within the cell but also how they are all coordinated. The rest of this section will describe the major biochemical processes in three main compartments of the plant cell: chloroplast, cytosol and mitochondria, that are of relevance to this thesis. An overview of the interaction between these compartments is provided in Section 2.2.4.

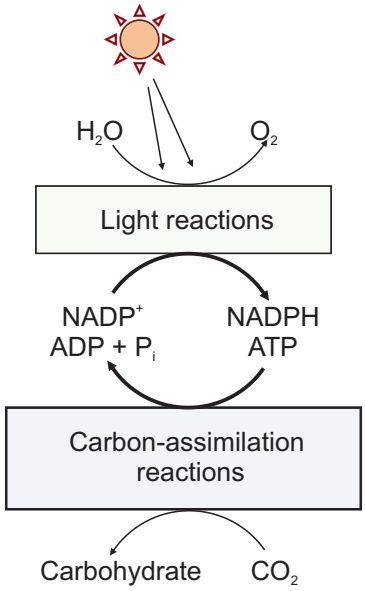


Figure 2.2 – An overview of the photosynthetic light reactions and carbon-assimilation reactions. See text for detailed description and List of Abbreviations for metabolite abbreviations. Adapted from [109].

2.2.1 Metabolic reactions of the chloroplast

Apart from the ancient chemolithotrophs[2], the profusion of life on Earth is supported entirely by radiant energy from the sun. The process that captures the energy from sunlight and converts it into the chemical energy of reduced inorganic compounds for use by all living organisms is called *photosynthesis*. Photosynthesis in plants encompasses two processes: the light reactions and the carbon-assimilation reactions. Light reactions occur only when plants are illuminated and convert solar energy to chemical energy in the form of ATP and NADPH. Carbon-assimilation reactions (also called dark reactions), on the other hand, use the ATP and NADPH produced during light reactions to reduce atmospheric carbon dioxide (CO_2) to form carbohydrates (such as glucose).

In plants and algae, both the light reactions and the carbon-assimilation reactions take place in specialised cytoplasmic compartments called chloroplasts. They contain an outer membrane that is permeable to small molecules and ions and a selectively-permeable inner membrane that encloses the internal compartment [110]. This compartment contains many flattened vesicles, called *thylakoids*, arranged in stacks called *grana*. The energy-transducing membranes of thylakoids contain the light capturing machinery of the chloroplasts, the *chlorophylls* and accessory pigments (e.g. carotenoids). These proteins, together with specific enzyme complexes embedded in the thylakoid membrane, carry out the light reactions to produce ATP and NADPH. The liquid space within which the thylakoids are embedded, called the *stroma*, contain enzymes that can use this ATP and NADPH to assimilate carbon from the atmosphere.

[2] Organisms that derive energy from oxidation of inorganic compounds.

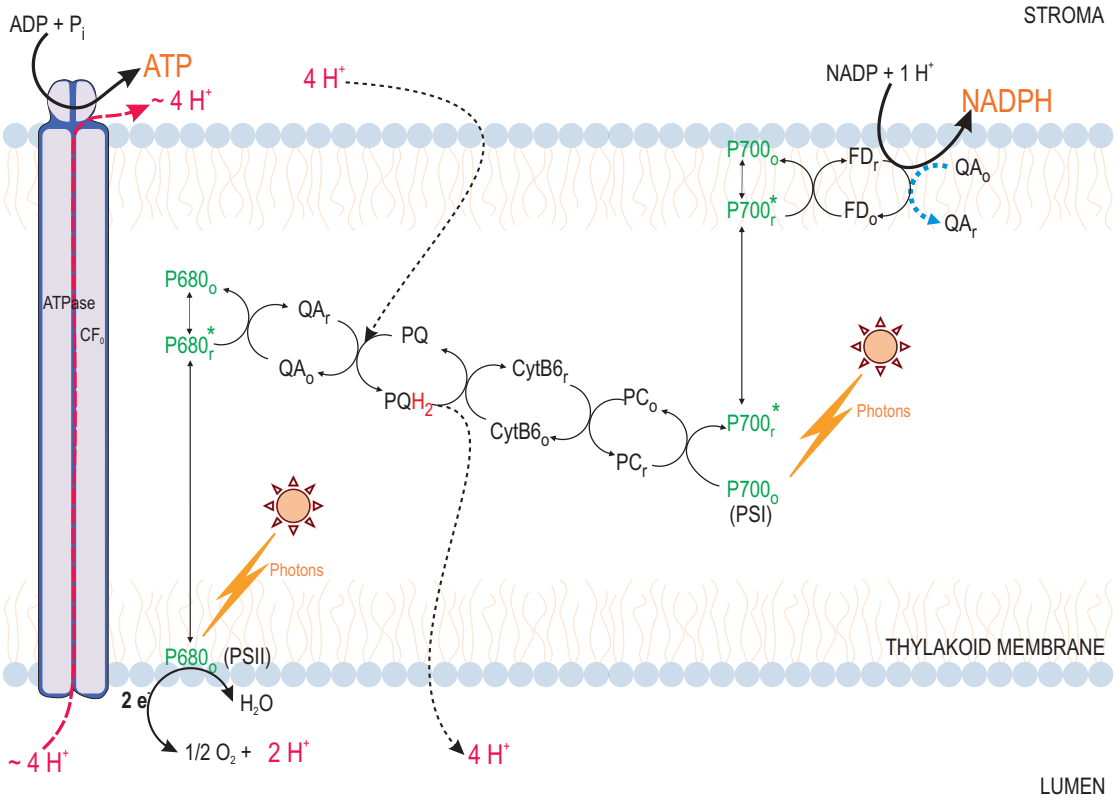


Figure 2.3 – Photosynthetic electron transport chain. Some electron carriers are not shown for clarity. See text for detailed description and List of Abbreviations for metabolite abbreviations.

2.2.1.1 Light reactions

The light-absorbing pigments of the thylakoid membranes are arranged in light harvesting units called *photosystems*. Although all of the pigment molecules (chlorophylls and carotenoids) in a photosystem are capable of absorbing photons, only a few chlorophyll molecules associated with the *reaction centre* are involved in transducing light into chemical energy. The former, called the light-harvesting pigments, act by transmitting the absorbed light energy rapidly (via other chlorophyll molecules) to molecules at the reaction centre. These chlorophyll molecules, consequently, become excited and release an electron (oxidation) to the nearby electron acceptor (reduction)s. This leaves the reaction centre chlorophyll with a missing electron, which it replaces with an electron from an electron-donor molecule. This way, excitation of chlorophyll molecules by light result in the initiation of an oxidation-reduction chain, commonly referred to as the electron transport chain (ETC).

The thylakoid membranes of the chloroplast have two different kinds of reaction centres, namely *photosystem I* (PSI) and *photosystem II* (PSII). These two reaction centres act in tandem to produce ATP and NADPH during light reactions (Figure 2.3). Synthesis of ATP during photosynthesis is called *photophosphorylation* [111, 112]. Two types of photophosphorylation are observed in plants: non-cyclic and cyclic. During

non-cyclic photophosphorylation, the reaction centre designated P680 in PSII absorbs photons, becomes excited and produces $P680^*$. Being an excellent electron donor, it donates its electrons to pheophytin acceptors, giving it a negative charge. The electron that was lost from P680 is regained by taking up the electron released during the oxidation of water inside the lumen (aqueous phase inside the thylakoids).

Further down the ETC, electrons reduce plastoquinone (PQ) to PQH_2. PQ takes up two protons from the stroma on reduction and pumps them into the lumen when oxidised. Eventually, the electrons in PQH_2 pass through cytochrome b_6f (CytB6) complex and reach the reaction centre P700 in PSI. The photochemical events that follow excitation of P700 are similar to those following excitation of P680. One exception is that the electron acceptor in this case is ferredoxin (FD), a protein loosely associated with the thylakoid membrane. Reduced ferredoxin is oxidised by transferring the electrons to $NADP^+$ with the help of the enzyme ferredoxin:$NADP^+$ oxidoreductase to form NADPH (Figure 2.3). This process is called non-cyclic photophosphorylation as the electrons flow from PSII through PSI to $NADP^+$ and are not recycled [111].

Cyclic photophosphorylation, however, involves only PSI. Electrons flowing from P700 to ferredoxin do not reach $NADP^+$; instead they move back through the quinones (QA) to PQ (indicated by the blue dotted arrow attached to FD in Figure 2.3) . Reduction of PQ to PQH_2 pumps more protons into the lumen. Upon oxidation PQH_2 donates the electrons to PC, which then transfers them to P700. This cyclic electron flow occurs to varying degrees depending on the environmental conditions, primarily light. It neither oxidises water to evolve oxygen nor produces NADPH, but acts by pumping electrons into the lumen resulting in valuable ATP generation [113].

Protons generated during the oxidation of water and those that are pumped into the lumen during ETC, both cyclic and non-cyclic, create a transmembrane electrochemical gradient of hydrogen ion concentration and membrane potential known as the proton motive force (PMF) across the thylakoid membrane. PMF is coupled with membrane-intrinsic 'coupling factor' CF_0 of the ATPase protein complex. CF_0 in chloroplasts has 14 proton binding sites. One complete PMF induced rotation of CF_0 can pump 14 protons to the stroma, leading to the formation of three ATP. Two NADPH molecules are produced during a single non-cyclic photophosphorylation step. But the exact number of ATP produced is still under debate [112]. Latest reviews on this topic are centred on the requirement of ATP and NADPH for the carbon-assimilation reactions, i.e. in the ratio of 3/2 [112]. Because of this uncertainty, the role of cyclic photophosphorylation in photosynthesis is also debated. One argument is that the cyclic electron flow acts to compensate the ATP deficit in the plant cell during adverse environmental conditions [112, 114, 115].

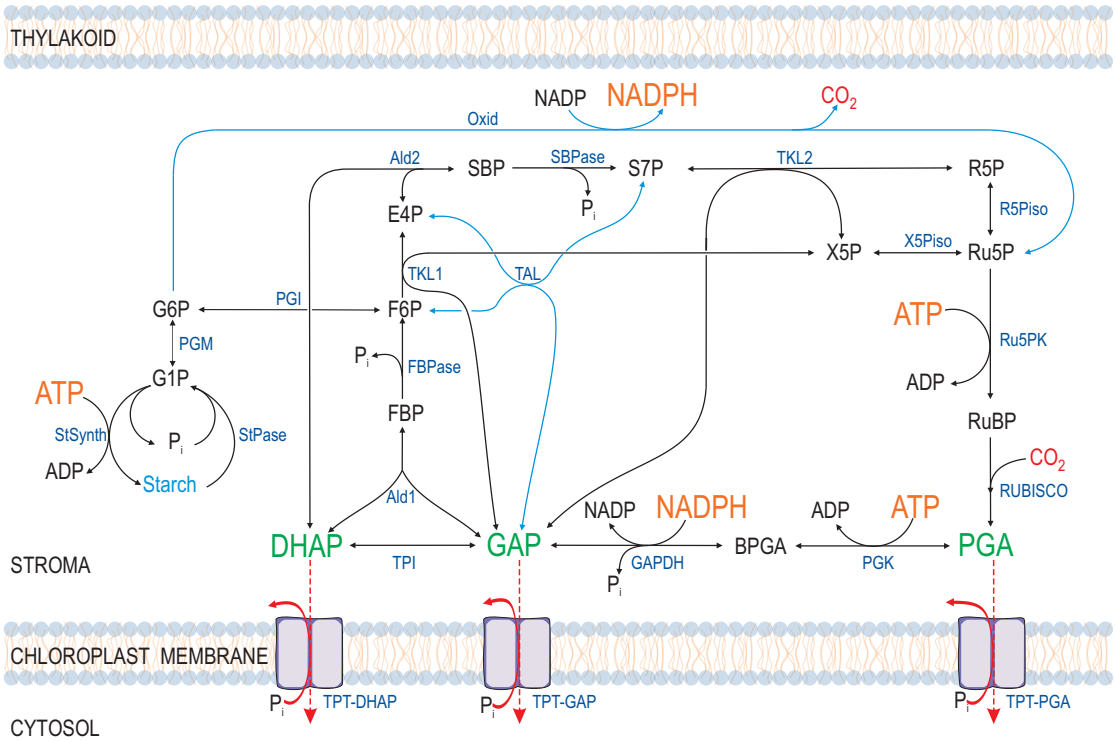


Figure 2.4 – Reactions of the Calvin cycle (black arrows) and oxidative pentose phosphate pathway (blue arrows). See text for detailed description and List of Abbreviations for metabolite abbreviations.

2.2.1.2 Carbon-assimilation reactions

The carbon-assimilation reactions in plants (and other autotrophs) synthesise carbohydrates from atmospheric CO_2 by reducing it at the expense of the ATP and NADPH produced during the light reactions. This process, also referred to as *CO_2 assimilation* or *carbon fixation*, takes place via a cyclic pathway[3] occurring within the chloroplast stroma. This pathway was elucidated in the early 1950s by Melvin Calvin and coworkers, and is often called the *Calvin cycle* (Figure 2.4).

The assimilation of CO_2 into biomolecules occurs in three stages: *carboxylation*, *reduction* and *regeneration* [116]. The first stage involves the incorporation of CO_2 and water into the five-carbon acceptor, ribulose 1,5-bisphosphate (RuBP), with the help of the enzyme ribulose 1,5-bisphosphate carboxylase oxidase (rubisco) to form two molecules (one of which contains the carbon atom from CO_2) of the three-carbon compound 3-phosphoglycerate (PGA). Rubisco is one of the most crucial enzyme in the production of biomass from CO_2. It accounts for almost 50% of the soluble proteins in chloroplasts [117]. Rubisco is not absolutely specific for CO_2 as a substrate. O_2 competes with CO_2 at the active site, and about once in every three or four turnovers, rubisco catalyses the condensation of O_2 with RuBP to form PGA and 2-phosphoglycolate [117]. This process, photorespiration, results in no fixation of

[3] Here, a pathway refers to a series of chemical reactions where the product of one enzyme-catalysed reaction is a substrate for another.

carbon, instead it consumes cellular energy and releases some previously fixed CO_2. 2-phosphoglycolate produced during photorespiration is converted to glycolate by a phosphatase and is exported to the peroxisome. Interested readers are directed to [117] for a detailed description of the glycolate pathway. Analysis of the metabolic pathways sequestered in peroxisomes is outside the scope of this thesis.

In the reduction phase, the PGA formed in the first stage is converted to glyceraldehyde 3-phosphate (GAP) in a two step process. The first step of which is catalysed by the enzyme 3-phosphoglycerate kinase (PGK) that attaches the phosphoryl group of an ATP molecule to PGA to form 1,3-bisphosphoglyceric acid (BPGA). BPGA is then reduced in the next step with the help of the electrons donated by a molecule of NADPH in a reaction catalysed by the enzyme glyceraldehyde 3-phosphate dehydrogenase (GAPDHP), yielding GAP.

In the carboxylation phase RuBP is consumed during the assimilation of CO_2. The regeneration phase contains a series of reactions that regenerate RuBP from GAP to ensure the continuous flow of CO_2 into carbohydrate. Figure 2.4 illustrates the various routes by which this is achieved. A reversible condensation of GAP with DHAP yields fructose 1,6-bisphosphate (FBP), the cleavage product of which, fructose 6-phosphate (F6P), is converted to starch by the enzymes in the stroma. Starch is temporarily stored in the chloroplast as insoluble granules. The rest of the excess GAP, DHAP and PGA are exported into the cytosol via specialised transporters to act as the precursors for sucrose synthesis and as the intermediates of glycolysis [118, 119]. Both these processes and the transport of Calvin cycle intermediates to the cytosol will be described elsewhere in this chapter.

An important mechanism responsible for coordinating changes in Calvin cycle enzyme activity in response to changes in light is the thioredoxin system [120, 116]. Thioredoxin is a small, mobile, disulphide-containing protein capable of accepting electrons moving from PSI through FD during the electron transport chain of the light reactions. Upon receiving the electrons, disulphide bonds in thioredoxin are reduced in a reaction catalysed by the enzyme ferredoxin-thioredoxin reductase. Reduced thioredoxin donates electrons for the reduction of the disulphide bonds of the four light-activated enzymes of the Calvin cycle: GAPDHP, fructose 1,6-bisphosphatase (FBPase), sedoheptulose 1,7-bisphosphatase (SBPase) and ribulose 5-phosphokinase (Ru5PK). These reductive reactions are accompanied by conformational changes that increase the activity of the enzymes [117]. A fifth enzyme rubisco is indirectly activated by the reduction of a related enzyme rubisco activase by thioredoxin [121]. In the absence of light, however, the disulphide bonds are re-oxidised and the enzymes are inactivated resulting in the cessation of CO_2 assimilation.

Another set of enzymes that is regulated by this light-driven reduction mechanism is part of a second pathway in the stroma, called the oxidative pentose phosphate pathway (OPPP), indicated by the blue arrows in Figure 2.4 [122]. The enzymes glucose 6-

phosphate dehydrogenase (G6PDH), 6-phosphogluconate dehydrogenase (6PGDH) and transaldolase of this pathway are up-regulated in the dark and down-regulated in the light [123, 124, 63]. These enzymes support the starch accumulated in the presence of light to be degraded to fuel glycolysis and sucrose synthesis at night (Figure 2.4). The major enzymes that are active at night include hydrolases (e.g. amylases and debranching enzymes), phosphorylases and glucanotransferases. The participation of all these enzymes lead to the hydrolysis of starch primarily to maltose and glucose, that are subsequently exported to the cytosol [125]. They are used for the synthesis of sucrose for export from the leaf, and for cellular metabolism.

2.2.2 Glycolysis

Glycolysis is the 'central' metabolic pathway of glucose catabolism that is ubiquitous, at least in part, in all organisms [123]. During glycolysis, a molecule of glucose is degraded in a series of enzyme-catalysed reactions to yield two molecules of the three-carbon compound pyruvate [126]. In plants, glycolysis furnishes the requisite metabolic options needed to facilitate growth and development by oxidising hexoses derived from starch and sucrose to generate ATP, reductant (NADH) and pyruvate. The latter two are taken up by the mitochondria to generate more energy equivalents and to support respiration (Section 2.2.3). Glycolysis is an amphibolic[4] pathway that can function in reverse to produce sucrose from low-molecular-weight compounds in an energy-dependent process referred to as gluconeogenesis [127, 126]. The sucrose thus formed is exported to non-photosynthetic parts of the plants to support their growth and development.

The breakdown of hexoses to three-carbon pyruvate during glycolysis occurs in two phases. During the first phase glucose released from sucrose or starch by the action of the enzymes invertase (Inv) or $\alpha-$ and $\beta-$amylases, respectively, is phosphorylated with the help of the phosphoryl group donated from ATP to form glucose 6-phosphate (G6P) (Figure 2.5). The enzyme phosphoglucose isomerase (PGI) acts on G6P by converting it to F6P, which is again phosphorylated by the ATP-dependent enzyme phosphofructokinase (PFK) to produce FBP. This is in turn broken down into GAP and DHAP (which is readily isomerised to a second molecule of GAP) to form the final products of the ATP-utilising preparatory phase. The energy gain comes in the second phase where the two molecules of GAP are oxidised and phosphorylated by inorganic phosphate to form two molecules of 1,3-bisphosphoglycerate (BPGA). This reaction, catalysed by the phosphorylating NAD-dependent GAP dehydrogenase (GAPDHP), is coupled to the formation of two NADH molecules. In the next step the two molecules of BPGA are converted to two molecules of pyruvate, leading to the formation of four

[4] A biochemical pathway that involves both catabolism and anabolism.

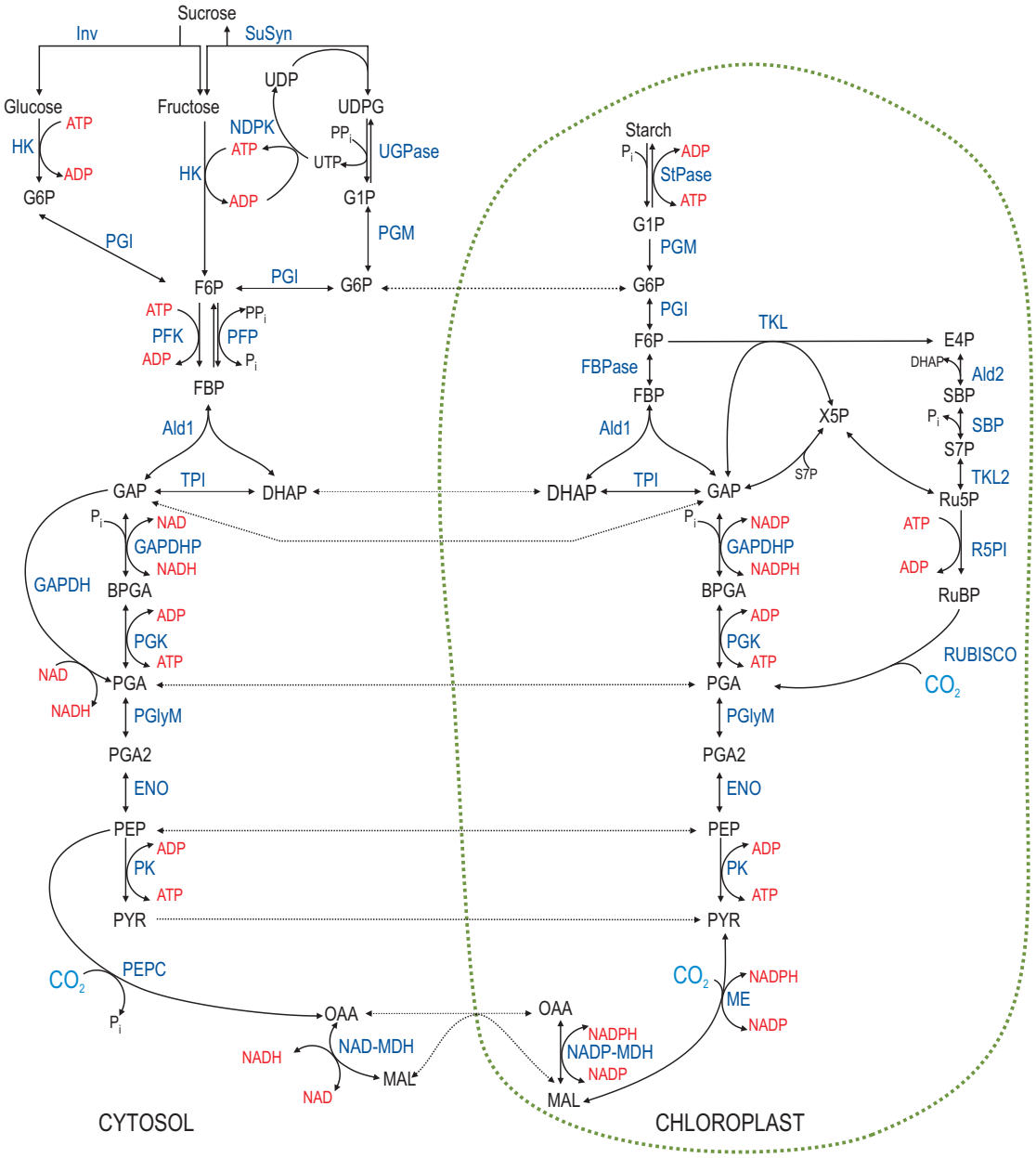


Figure 2.5 – Glycolytic reactions of the chloroplast (dotted green enclosure) and cytosol. Reactions of the Calvin cycle are also shown in the chloroplast. Transports and shuttle mechanisms of the inner chloroplast membrane are shown using dotted lines. See text for detailed description and List of Abbreviations for metabolite abbreviations.

molecules of ATP. This conversion is performed via sequential reactions catalysed by the enzymes PGK, phosphoglycerate mutase (PGlyM), enolase (ENO) and pyruvate kinase (PK). A detailed review of the various enzymes and their role in plant glycolysis is available in [127] and [128].

Glycolysis in plants can occur independently in two subcellular compartments, the cytosol and the chloroplasts. These parallel glycolytic reactions are (Figure 2.5) catalysed by isozymes[5] encoded by distinct nuclear genes [127]. The presence of isozymes in plants is attributed to the need for separate enzymes capable of catalysing

[5] Two or more enzymes that catalyse the same reaction but are encoded by different genes.

similar reactions in different subcellular compartments [129]. The genes coding for isozymes vary from each other through the changes in amino acid composition, which will often alter the charge or, in some cases, the physical and kinetic properties of the enzyme [130]. Such variations will adapt the isozymes for efficient catalysis in different compartments having specific metabolite concentrations and pH. Enzymes involved in plastid glycolysis are synthesised as inactive precursors in the cytosol. They are then imported into the chloroplast with concomitant cleavage of an *N*-terminal transit peptide [127].

Glycolysis in the cytosol and in the chloroplast differs with their specific roles. In cytosolic glycolysis, the major substrate, sucrose, is oxidised by the glycolytic enzymes in the cytosol to generate energy equivalents, reductant and building blocks for anabolism. In photosynthetic chloroplasts, however, glycolysis serves to convert the 'excess' intermediates of the Calvin cycle (GAP, DHAP and PGA) to pyruvate. In non-photosynthetic plastids and in chloroplasts in the dark, glycolytic enzymes participate in the breakdown of starch to triose-phosphate intermediates and finally to pyruvate [131, 118, 132]. Plastidic and cytosolic glycolysis can interact through the action of a number of highly selective transporters in the chloroplast membrane, the properties of which are described elsewhere in this chapter. These transporters mediate the transfer of intermediates of the Calvin cycle and plastidic glycolysis into the cytosol to fuel cytosolic glycolysis.

2.2.3 Mitochondrial metabolism

Mitochondria are pleomorphic organelles composed of a smooth outer membrane surrounding an inner membrane that has convolutions (cristae) designed to attain increased surface area. Unlike the outer membrane, the inner membrane is impermeable to most small molecules and ions, including protons. It surrounds a protein-rich core, called the matrix, containing DNA, ribosomes and enzymes particular to the mitochondria [123].

Specific transporters in the mitochondrial inner membrane carry the pyruvate produced in the cytosol during glycolysis into the matrix. Here, pyruvate is converted to acetyl-CoA (ACoA) and CO_2 in an irreversible oxidative decarboxylation reaction catalysed by the pyruvate dehydrogenase (PDH) complex containing three different enzymes. The combined oxidation and decarboxylation of pyruvate requires the sequential action of coenzyme-A (CoA-SH) and NAD^+ along with three other coenzymes[6] [133].

ACoA produced in the above reaction undergoes oxidation carried out by a series of reactions called the *tricarboxylic acid (TCA) cycle* (Figure 2.6). Initially, ACoA donates its acetyl group to the four-carbon compound oxaloacetate (OAA) to form the six-

[6] For a complete description of this reaction, please refer [133].

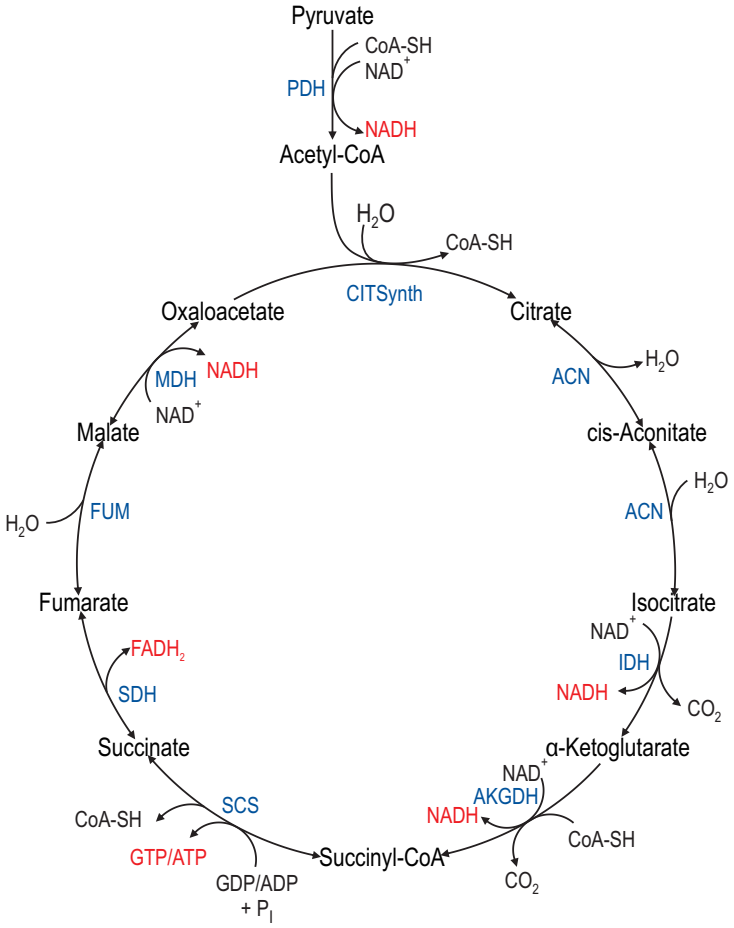


Figure 2.6 – Reactions of the tricarboxylic acid cycle. See text for detailed description and List of Abbreviations for metabolite abbreviations.

carbon citrate (CIT) in a condensation reaction catalysed by the enzyme citrate synthese (CITSynth). CIT is then converted to isocitrate (IsoCIT), another six-carbon compound, in the subsequent sequential dehydration and hydration reaction catalysed by the enzyme aconitase (ACN). Ensuing oxidative decarboxylation reaction catalysed by isocitrate dehydrogenase (IDH) produces five-carbon α-ketoglutarate (AKG), NADH and CO_2 from IsoCIT. The former product is converted to succinyl-CoA (SCoA) in a similar reaction catalysed by the α-ketoglutarate dehydrogenase (AKGDH) enzyme complex. This reaction takes up CoA-SH and releases NADH and CO_2. In the next step, SCoA is broken down to the four-carbon succinate (SUC) by the enzyme succinyl-CoA synthetase (SCS). Energy released from the breakdown of the thioester (-SH) bond in SCoA during this reaction is used to drive the synthesis of a molecule of GTP or ATP. Oxidation of SUC is carried out by the flavoprotein succinate dehydrogenase (SDH). Formation of the end product of this reaction, fumarate (FUM), is thus coupled to the reduction of a molecule of ubiquinone (Q) to form QH_2 and an ensuing electron transport chain (see next paragraph). $FADH_2$ produced during this process is considered as a molecule of NADH in this thesis. FUM is then enzymatically converted in two steps into the four-carbon OAA which is ready to react with another molecule of ACoA. The

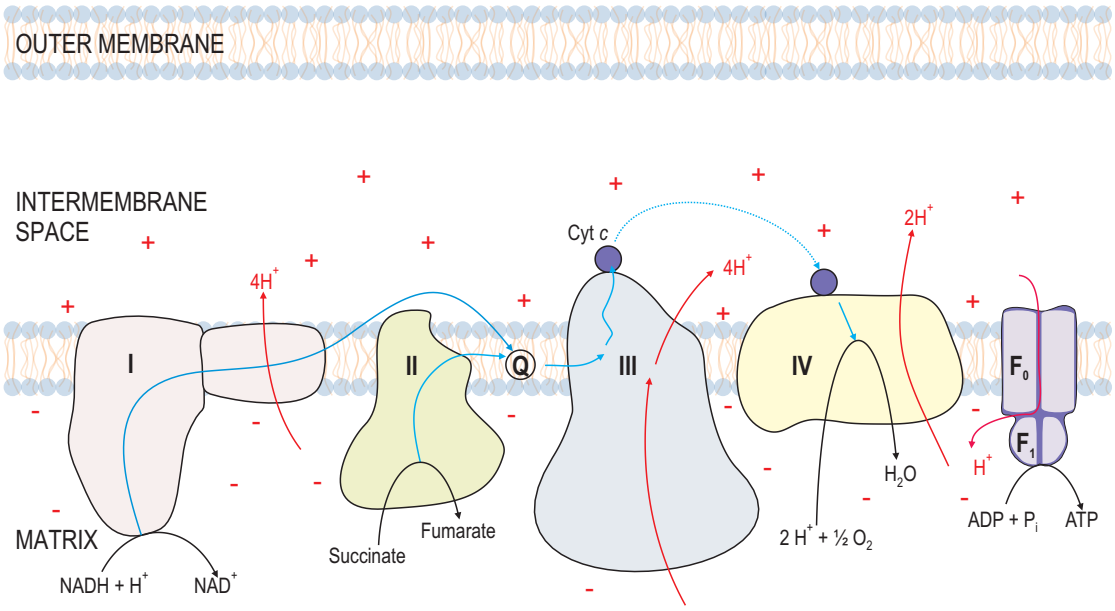


Figure 2.7 – The mitochondrial electron transport chain. Blue arrows indicate the direction of the flow of electrons. '+' and '-' indicate positive and negative proton gradient, respectively. See text for detailed description and List of Abbreviations for metabolite abbreviations. Adapted in part from [109].

hydration and dehydration steps in this reaction are accompanied by the release of a molecule of NADH.

In each turn of the TCA cycle four molecules of NADH are produced from a molecule of pyruvate. Oxidation of these coenzymes release electrons that are carried through electron carriers to form ATP and water in a process referred to as *oxidative phosphorylation* (Figure 2.7). These electron carriers of the mitochondria are organised into molecular complexes embedded in the inner membrane. Complex I (NADH dehydrogenase) and II (succinate dehydrogenase) catalyse the electron transfer to Q from two different electron donors: NADH (Complex I) and succinate (Complex II). Reduced Q (QH_2) serves as a carrier of electrons. It passes the electrons to Complex III, which then passes them to another carrier, cytochrome *c*. Cytochrome *c* delivers the electrons to Complex IV, which completes the sequence by transferring them from reduced cytochrome *c* to O_2 [109]. Electron flow through Complexes I, III and IV is accompanied by the flow of protons from the matrix to the intermembrane space (IMS), resulting in both a chemical gradient (ΔpH) and an electrical gradient ($\Delta\psi$) across the membrane. As the inner mitochondrial membrane is impermeable to protons, they can re-enter the matrix only through the proton-specific F_0 channel of the ATPase complex. The PMF generated during the movement of protons back into the matrix through this channel results in ATP synthesis, catalysed by the F_1 complex associated with F_0 [134, 109]. This process is exactly analogous to photophosphorylation described in Section 2.2.1.1.

ATP synthesised in the mitochondria are exported to the cytosol via specialised antiporters referred to as adenosine nucleotide translocases or ATP/ADP translocators in exchange for cytosolic ADP. In non-photosynthetic eukaryotes and photosynthetic eukaryotes in the dark, mitochondria are the site for most energy-yielding oxidation reactions and ATP synthesis. In photosynthetic eukaryotes in the presence of light, however, chloroplasts produce most of an organism's ATP [135].

Although the central role of mitochondria in plant cells is to carry out energy-yielding metabolism, it is also a site for the synthesis of vitamin cofactors and amino acids [135]. The majority of these end products are synthesised from the four- and five-carbon intermediates of the TCA cycle. Isolated mitochondria have been found to contain the genome and protein-synthesising machinery required for synthesising mainly electron transport proteins and parts of ATP synthase [136]. The majority of the mitochondrial polypeptides are encoded in the nuclear genome, synthesised in the cytosol and imported into the mitochondria via specific transporters [137, 138].

2.2.4 Interaction between compartments

The integration of cellular metabolism necessitates interaction between the metabolic pathways sequestered in various subcellular compartments. Metabolism in every compartment depends on these interactions for supplies of energy (ATP), redox equivalents and metabolic precursors. This section will provide an overview of such interactions between the compartments — chloroplasts, cytosol and mitochondria — a summary of which is shown in Figure 2.8. Note that the scope of this thesis allows the description of only those aspects of the interaction that are of relevance to the modelling and analysis described elsewhere.

2.2.4.1 Metabolite exchange between chloroplast and cytosol

The selectively-permeable chloroplast inner membrane represents the interface between chloroplast and cytosol [110]. It contains a variety of transporters that mediate the exchange of metabolites between both compartments [139, 140, 141, 132, 142, 143]. In the presence of sunlight, CO_2 fixed by the Calvin cycle is exported from the chloroplast into the cytosol for the synthesis of sucrose and pyruvate. While the former is allocated to the heterotrophic organs of the plant such as roots, seeds or fruits, the latter is either used for amino acid synthesis or undergoes further oxidation in the mitochondria to produce more energy and reductant. Export of the newly-fixed carbon in the form of the intermediates of the Calvin cycle, triose-phosphates (GAP and DHAP) and PGA, is mediated by the triose-phosphate/phosphate translocators (TPT) of the chloroplast inner membrane. These proteins catalyse a strict 1:1 counter-exchange of triose-phosphates or PGA in the stroma with inorganic orthophosphate (P_i) in the cytosol [118]. The import

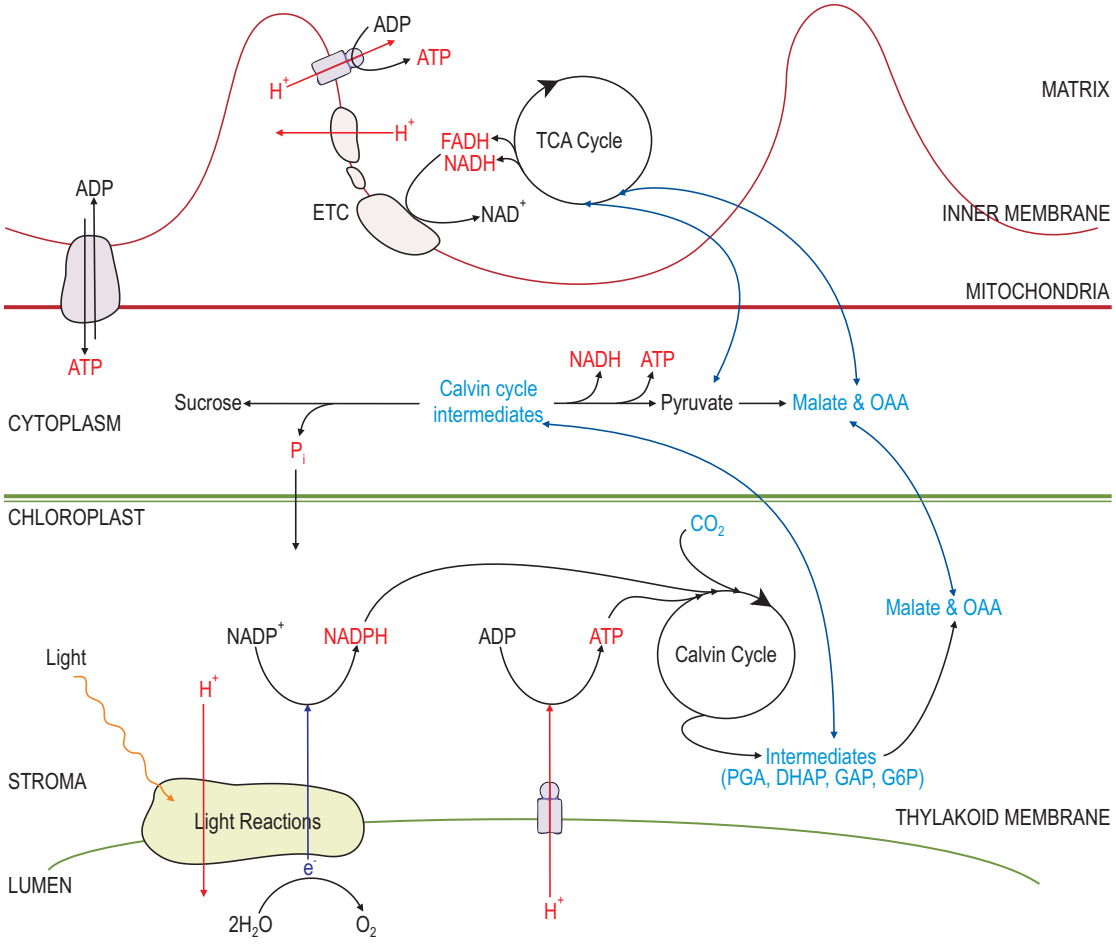


Figure 2.8 – Summary of metabolite interactions between the metabolic pathways sequestered in chloroplast, cytosol and mitochondria. See text for detailed description and List of Abbreviations for metabolite abbreviations.

of P_i ensures that the export and the uptake of the phosphate moieties are balanced, and allows continuous ATP synthesis.

Exchange of another intermediate of the Calvin cycle, G6P, is mediated by the glucose 6-phosphate/phosphate translocator (GPT). The main function of GPT is the import of G6P into non-photosynthetic plastids and chloroplasts in the dark for use as a precursor for the synthesis of starch and the OPPP. Such plastids are devoid of FBPase activity, the key enzyme for the conversion of triose-phosphates to G6P. Therefore, they rely on the supply of G6P from the cytosol via GPT. In contrast, FBPase is highly active in photosynthetic chloroplasts. G6P produced during this reaction is often exchanged with the cytosol via GPT [141, 143].

Another important phosphate translocator of the chloroplast membrane is the phosphoenolpyruvate/phosphate translocator (PPT). Some chloroplasts do not possess the enzymes required for the conversion of PGA to PEP [144, 143]. PEP, however, is strictly required for various plastid-localised metabolic pathways including shikimate pathway and the biosynthesis of various fatty acids and amino acids. Therefore, the proposed function of PPT in such chloroplasts is the transport of PEP into the

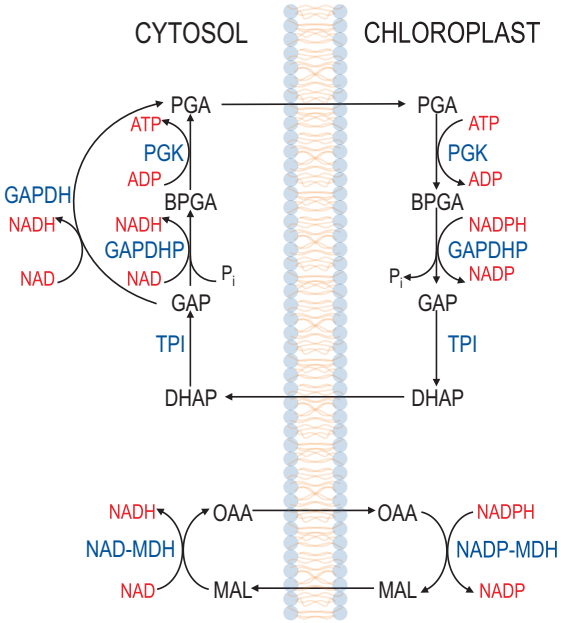


Figure 2.9 – Scheme of the indirect transfer of ATP and redox equivalents between chloroplast and mitochondria. See text for detailed description and List of Abbreviations for metabolite abbreviations.

organelle [145]. In other chloroplasts, PEP is exported to the cytosol, where it is required by the enzyme phosphoenolpyruvate carboxylase (PEPC) for primary carbon fixation [143]. The characteristics of PEP exchange via PPT are similar to those of TPT [146, 141].

Besides phosphate translocators, the chloroplast inner membrane harbours other important transport proteins, such as those responsible for the exchange of malate (MAL) and OAA. OAA in the chloroplast is used to produce aspartate, a precursor for the biosynthesis of various amino acids. It can also be reduced to MAL in a reaction catalysed by the stromal NADP-dependent enzyme MAL dehydrogenase (NADP-MDH). In the cytosol, MAL produced in a similar reaction is catalysed by NAD dependent MDH (NAD-MDH). Malate in the chloroplast is exported to the cytosol via MAL transporters that are linked to the OAA transporters. Any import of MAL through these transporters results in a concomitant export of OAA, resulting in a MAL/OAA antiport, commonly referred to as the MAL/OAA shuttle [144].

A very important non-phosphate translocator of the chloroplast inner envelope is the ATP/ADP translocator. The activity and significance of this transport protein in the interaction between various subcellular compartments will be discussed in the following sections.

2.2.4.2 Transfer of redox equivalents and ATP between chloroplast and cytosol

As outlined previously, the photosynthetic ETC in the chloroplast converts light energy to chemical energy in the form of redox equivalents (NADPH) and ATP, which are

used by the Calvin cycle to fix atmospheric carbon. The excess NADPH and ATP produced during the light reactions contribute to meet the cell's energy demands for growth and maintenance, and are exported to the cytosol so that they are accessible for other organelles. Neither of these molecules, however, can penetrate the selectively-permeable inner chloroplast membrane because of their charge and size.

Nevertheless, the transfer of redox equivalents from the chloroplast to the cytosol can occur indirectly by the MAL/OAA shuttle or by the DHAP/PGA shuttle (Figure 2.9) [147, 148, 149]. The operation of these shuttle mechanisms is driven by the existence of a redox gradient between chloroplast and cytosol. Experiments conducted using a non-aqueous fractionation procedure showed that the ratio of NADPH/NADP in the stroma of illuminated spinach leaves is substantially higher than the NADH/NAD ratio in the cytosol [148]. In MAL/OAA shuttle, the redox transfer is controlled by a light inducible $NADP^+$-malate dehydrogenase (NADP-MDH) enzyme that is activated when the stromal NADPH/NADP ratio is very high [150, 148]. Under such conditions OAA is converted to MAL and subsequently exported to the cytosol via the MAL/OAA transporter located in the inner chloroplast membrane. In this way, the excess photosynthetic redox equivalents are released to the cytosol and the redox gradient between the stromal NADPH/NADP and cytosolic NADH/NAD is alleviated.

With the DHAP/PGA shuttle, the provision of redox equivalents in the form of NADH is accompanied by the export of ATP. In this shuttle, catalysed by the triose-phosphate translocators, the export of stromal DHAP is coupled to the import of cytosolic PGA. Once in the cytosol, DHAP is oxidised to PGA, which is then ready to be imported into the stroma to finish the cycle (Figure 2.9). For every molecule of DHAP exported, one molecule of both ATP and NADH, or a molecule of NADH alone, is liberated into the cytosol depending on the cytosolic enzymes involved (no ATP is produced when cytosolic DHAP is converted to PGA by the NAD-dependent non-phosphorylating GAPDH, which produces NADH only) [148, 151]. In this way, ATP and NADPH consumed during the reduction of PGA to DHAP in the chloroplast are released into the cytosol during the oxidation of DHAP to PGA. The redox gradient between the NADPH/NADP ratio in the stroma and the NADH/NAD ratio in the cytosol is maintained by this shuttle, primarily by limiting the oxidation of DHAP to PGA in the cytosol [148, 152].

Although ATP/ADP translocators are present on the selectively-permeable inner membrane of the chloroplast [153, 154, 147, 142], the export of ATP during the DHAP/PGA shuttle has been found to be a more efficient means by which photosynthetic ATP can be exported from chloroplast to cytosol [149]. Unlike the highly-active ATP/ADP transporters on the mitochondrial membrane that can rapidly export ATP from the matrix to the cytosol in exchange for ADP [139, 140], the activity

and affinity of those on the chloroplast membrane are very low and suited only for importing ATP into the chloroplast [154, 140, 144]. The maximum rate of this ATP import, generally observed in young leaves, has been found to be 10-fold lower than that of other metabolites such as P_i [140, 155], and hence it is considered unlikely to play any significant role in photosynthetic leaf cells [156]. However, in non-photosynthetic plastids and chloroplasts at night import of energy in the form of ATP is deemed necessary to energise anabolic and catabolic reactions such as starch and fatty acid synthesis in storage plastids [157], and starch degradation in chloroplasts at night [158]. The import of ATP into the stroma is also regulated by the distribution of ATP/ADP gradients in the cell. In photosynthesising and non-photosynthesising leaves, the ATP/ADP ratio is higher in the cytosol than in the stroma [148].

2.2.4.3 Energy and metabolite exchange between cytosol and mitochondria

Like chloroplasts, mitochondria can also export ATP and redox equivalents to the cytosol. The highly-active ATP/ADP transporters on the mitochondrial membrane mediate a direct transfer of ATP to the cytosol with concomitant import of ADP. As the ATP/ADP ratio outside the mitochondria is much higher than that in the matrix [159], the proton gradient generated during the mitochondrial ETC is required for transferring the newly-formed ATP to the higher phosphorylation potential of the cytosolic ATP/ADP system [160].

Redox equivalents formed in the mitochondrial matrix can be either re-oxidised in the ETC or exported to the cytosol. In contrast to the ATP/ADP transport, the export of redox equivalents is not direct; it is mediated by several metabolite shuttles catalysed by mitochondrial membrane transport proteins [139, 151, 149, 152]. An example for such shuttle mechanisms is the CIT/MAL exchange, where the subsequent decarboxylation of citrate exported into the cytosol to 2-oxoglutarate (2-OG) results in the production of cytosolic NADH. The MAL/aspartate (Asp) exchange involving MAL/2-OG and glutamate/Asp translocators is another example. However, the contribution of these shuttle mechanisms to the transfer of reducing equivalents from mitochondria to the cytosol is minor when compared to the MAL/OAA shuttle [161]. In mitochondria, MAL and OAA are intermediates of the TCA cycle and their interconversion is mediated by the mitochondrial NAD-MDH (Figure 2.6). The export of MAL from mitochondria is driven by the cytosolic NADH/NAD ratio, which, in the presence of light, is 70 times lower than in the mitochondrial matrix [151].

2.2.4.4 Interaction between chloroplast and mitochondria

Although chloroplasts and mitochondria are organelles separated by independent membrane envelopes and the cytosol in between, they are not only interdependent in

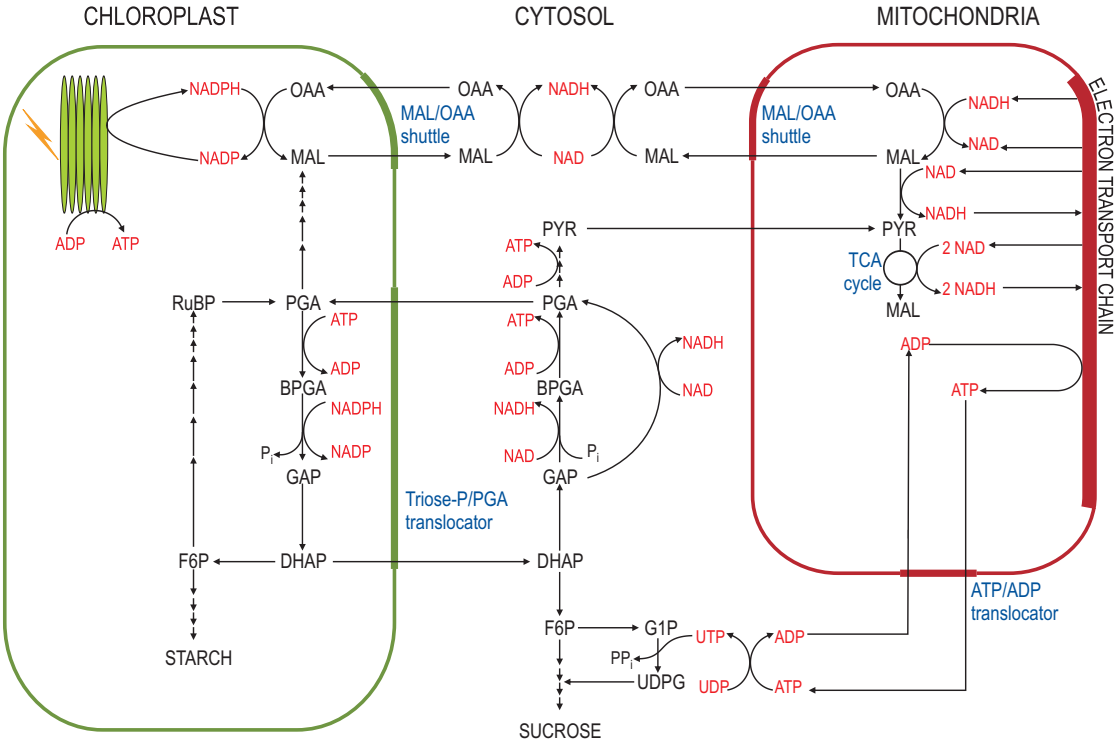


Figure 2.10 – Summary of ATP and redox equivalent interactions between the metabolic pathways sequestered in chloroplast, cytosol and mitochondria. See text for detailed description and List of Abbreviations for metabolite abbreviations.

their functions but also are mutually beneficial in their interaction. Several reviews exist that examine in detail the different aspects of the interaction between chloroplast and mitochondrial metabolism [151, 149, 152, 162, 163]. Interactions that involve the metabolic pathways and transport mechanisms described in the previous sections are discussed below.

Mitochondria are the main source of ATP for photosynthetic cells in the dark (Section 2.2.4.2). In the presence of light, however, the extent and efficiency of mitochondrial metabolism is still not well established [149, 162]. Excess ATP generated during photosynthesis is exported to the cytosol via the DHAP/PGA shuttle (Section 2.2.4.2). Once in the cytosol the adenylates can restrict respiration in various ways, for example, a high ATP/ADP ratio (greater than 20) can restrict respiration by reducing the concentration of ADP in the cytosol [149]. In spite of these restriction mechanisms, respiration continues in the presence of light to contribute ATP for energy-demanding processes such as nitrogen assimilation and adapting to environmental factors [149, 162, 163].

In photosynthesising chloroplasts, the components of the ETC become highly reduced when the NADPH/NADP ratio becomes too high. Over-reduction results in the photoinhibition[7] of chlorophyll molecules leading to reduced photosynthetic

[7] Photoinhibition is the reduction in a plant's capacity for photosynthesis due to the build-up of excess redox equivalents under various stress conditions such as excess light levels or suboptimal CO_2 or O_2

efficiency. Mitochondrial respiration can prevent over-reduction of ETC by oxidising the excess reductants generated in the chloroplast [149, 165]. Chloroplasts export reductants to the cytosol through the DHAP/PGA shuttle and the MAL/OAA shuttle (Section 2.2.4.2). MAL exported to the cytosol as part of the latter shuttle mechanism enters mitochondria through the mitochondrial OAA translocator, to be oxidised to OAA by mitochondrial NAD-MDH. NADH released during this process is either used for the production of ATP or exported to the cytosol via various mitochondrial shuttle mechanisms (Section 2.2.4.3). These redox equivalents are partly dissipated by the mitochondria during both carbon and ammonia assimilation [149, 162, 163]. NADH transferred to the cytosol is used for hydroxypyruvate reduction occurring as part of the photosrespiratory pathway in the peroxisomes, and nitrate reduction proceeding as a partial step of nitrate assimilation in the cytosol [155]. Similarly, ATP generated in the mitochondrial matrix is translocated to the cytosol via ATP/ADP translocators to be used in sucrose synthesis.

2.3 Models of plant metabolism

The use of computers to construct and analyse mathematical models of various aspects of plant metabolism has a history dating back at least as far as the late 1950s, with the work of Chance *et al.* published in 1960 [166]. They constructed a model of plant central metabolism containing the reactions involved in glucose phosphorylation, ATP utilisation, and glycolytic and oxidative phosphorylation to study the kinetics of intermediates such as ATP, ADP and glucose in a multi-component enzyme system. In a subsequent study, this model was extended by expanding some of the lumped reactions in each of the four components and by including reversible reactions [167]. The major constraints faced by the authors during these studies were the limitations of the available computational memory and power, and the lack of sufficient experimental data required for accurately defining the model [167].

During the mid 1970's, there were significant developments in computer technology that reduced the above limitations. A similar phase of progress was also seen in metabolic engineering and metabolic control analysis[8] (MCA) [168], mainly due to the emergence of a number of theoretical tools and concepts. The result was an array of mathematical models trying to capture the dynamic nature of plant metabolism, especially photosynthesis. First among them was a mathematical model of photosynthesis and photorespiration published in a Russian article by Laisk in 1973 [169]. Mathematical modelling of photosynthesis and estimation of parameters

levels [164].

[8] It is a method for analysing how the control of fluxes and intermediate concentrations in a metabolic pathway is distributed among the different enzymes that constitute the pathway.

from empirical data were then new approaches. Laisk's model was soon followed by the models of Thornley (1974) [170], Milstein and Bremermann (1979) [171] and Kaitala *et al.* (1982) [172]. The Thornley model (1974) was based on a novel approach in which a very simple mathematical model was formulated to represent the most important features of the dynamics of photosynthesis. This approach demonstrated the concept of constructing simple models that are fairly easy to analyse and interpret in order to study complex behaviours of the system. In contrast, Milstein and Bremermann (1979) studied a large complicated model containing 17 nonlinear ordinary differential equations (ODEs) to study the kinetic parameters of the Calvin photosynthesis cycle. Kaitala *et al.* (1982) used the approach of Thornley to construct a kinetic model of photosynthesis describing the effect of radiant energy and CO_2 concentration in the control of CO_2 assimilation in leaves. Thornley's approach is used elsewhere in this thesis to simplify a structural model of photosynthesis.

An important study that relates to this thesis was performed by Giersch *et al.* in 1980 [173] where the available data on phosphate translocators and the triose-phosphate oxidation system of the chloroplast envelope were used to construct a kinetic model to estimate the efficiency of indirect ATP transfer between chloroplasts and cytosol of leaf cells. They showed that the triose-phosphate/PGA shuttle is adequate to provide photosynthetic ATP for cytosolic reactions at physiologically meaningful rates. Their model also demonstrated the necessity of a transmembrane proton gradient for efficient indirect ATP transfer across the chloroplast envelope. However, this model did not take into account either the activity of MAL/OAA shuttle or the possibility of other routes through which redox equivalents and ATP can be transferred from the chloroplasts to cytosol. Similar studies employing models of plant metabolism, especially those of photosynthesis, were undertaken by Woodrow (1986) [174], Petterson and Ryde Petterson (1988) [175], Laisk *et al.* (1989) [176] and Giersch *et al.* (1991) [177]. Many more examples exist, a full review of which is outside the scope of this thesis. However, a few recent models that have some bearing to the work in this thesis are described in the rest of this section.

One major obstacle in modelling plant systems in the later half of the 1990's was the limitation in the availability of sufficient and reliable data pertaining to the activity of enzymes and the compartmentation of metabolites and reactions. However, rapid advances in high-throughput molecular biology techniques and the emergence of a number of theoretical concepts in metabolic modelling, both kinetic and structural, aided in reducing this limitation to some extent. Another obstacle were the constraints associated with computational memory and power. But this limitation was fast diminishing as developments in the field of computer technology continued to rise exponentially during this period. Along with computer memory and power, operating systems and user interfaces improved beyond recognition and a plethora of software for

Table 2.2 – Stoichiometries of elementary modes of Calvin cycle model in light (Adapted from [63]). 'Starch' is interpreted as one glucose unit arising from stromal starch. The subscript 'cyt' indicates a cytosolic metabolite. The last elementary mode in the table is a futile cycle comprising starch synthase and starch phosphorylase driven by ATP from the light reaction. See List of Abbreviations for metabolite abbreviations.

Substrate(s)	Product
3 CO_2	PGA_{cyt}
3 CO_2	$DHAP_{cyt}$
3 CO_2	GAP_{cyt}
3 CO_2 + Starch	3 PGA_{cyt}
3 CO_2 + Starch	3 $DHAP_{cyt}$
3 CO_2 + Starch	3 GAP_{cyt}
6 CO_2	Starch
Starch	Starch

undertaking metabolic studies was introduced. Armed with these advanced technologies and theoretical concepts, numerous attempts were made to construct sophisticated models of plant metabolism.

A kinetic model of the Calvin cycle was formulated to study its behaviour in the presence and absence of light (Figure 2.4) [178]. The model had 23 reactions and nearly as many metabolites representing the conversion of triose-phosphate intermediates to starch in the presence of light. The export of excess triose-phosphates from the chloroplast to the cytosol with concomitant import of cytosolic inorganic phosphate was represented using three transport reactions. The energy produced during photosynthetic light reactions was represented in the model using a single reaction producing stromal ATP. Light control on the model was implemented by including the thioredoxin system where FBPase, SBPase, Ru5PK, GAPDHP and rubisco are up-regulated (included during model analysis) in the light and down-regulated (removed during analysis) in the dark. Starch degradation and the subsequent metabolism in the dark were represented in the model by including the reactions of the OPPP that are down-regulated in the light and up-regulated in the dark (Figure 2.4). The resulting model was found to exhibit alternate steady-states of low or high carbon assimilation flux, with hysteresis in the transitions between the steady states induced by environmental factors such as phosphate and light intensity [179]. Further studies on this model revealed the existence of two separate steady-states in the photosynthetic Calvin cycle [180].

An interesting observation from the simulation of the above model is that at high cytosolic inorganic phosphate levels the total exported carbon flux via the phosphate translocators exceeds the assimilation flux via rubisco, corresponding to a situation in which the excess carbon requirement is being fulfilled by starch degradation [179]. This raises the question whether it is possible for starch degradation to enhance photosynthetic triose-phosphate export in the light. To answer this question a structural analysis involving EM analysis was performed on the Calvin cycle model that was described previously [63]. It was shown that all EMs of the system shown in Table 2.2, including

Table 2.3 – Overall stoichiometries of EMs in the dark (Adapted from [63]). Those metabolites that are subscripted 'ext' are cytosolic metabolites that have a stromal counterpart. See List of Abbreviations for metabolite abbreviations.

Substrate(s)	Product
Starch + P_{iext}	$G6P_{ext}$
Starch + P_{iext} + 2 NADP	$R5P_{ext}$ + 2 NADPH + CO_2
Starch + P_{iext} + 4 NADP	$E4P_{ext}$ + 4 NADPH + 2 CO_2
Starch + P_{iext} + 6 NADP	GAP_{ext} + 6 NADPH + 3 CO_2
Starch + P_{iext} + 6 NADP	$DHAP_{ext}$ + 6 NADPH + 3 CO_2
Starch + 12 NADP	12 NADPH + 6 CO_2

the starch degrading modes, have both rubisco and light reactions as components. This implies that none of these EMs can sustain flux in the absence of light and there is, therefore, an obvious link between starch degradation and triose-phosphate export.

Subsequent elementary modes analysis of the Calvin cycle model in the dark reiterated the aforementioned fact that the model is not able to perform starch degradation in the absence of other light-activated reactions (Table 2.3). It was shown that plants overcome this limitation by the inclusion of the OPPP and the thioredoxin system, which ensures the availability of sugar phosphates and NADPH in the dark by degrading starch accumulated during the light. EM analysis of this model also emphasised that there must be a tight coupling of triose-phosphate export to the reduction of NADP to NADPH, as in the absence of such a coupling the oxidative reactions of the OPPP would rapidly exhaust their supply of cosubstrate, NADP. However, to study this further the Calvin cycle model has to be extended by incorporating light reactions, nucleotide synthesis, the shikimate pathway and the redox exchange via shuttle mechanisms [63].

A similar study was performed on a model of carbohydrate metabolism in potato tuber cells [181]. This model had 29 reactions representing cytosolic sucrose synthesis and glycolysis, and starch degradation in amyloplasts. An important aspect that distinguished this model from previous efforts was the compartmentation of reactions and metabolites. Transfer of metabolites and ATP between the two compartments was represented using specific transport reactions. EM analysis was performed on this model to identify modes with the highest ATP yield and substantial starch and sucrose turnover.

Part II

Modelling

CHAPTER 3

Modelling plant carbon metabolism

3.1 Overview

The molecular and biochemical background of stromal, cytosolic and mitochondrial carbon metabolism in plant cells were reviewed in the previous chapter with special attention to the exchange of ATP and reducing equivalents between them. Furthermore, the basic techniques involved in the construction and analysis of structural metabolic models were described in Chapter 1. A major objective of this study is to investigate the characteristics of the interaction between metabolic pathways sequestered in chloroplasts, the cytosol and mitochondria using structural modelling and analysis techniques. Defining a metabolic model to study these interactions, however, necessitates manually reconstructing independent steady-state metabolic models representing pathways in each of these compartments and finally integrating them using specific transport reactions. For this reason, this chapter is devoted to the reconstruction and analysis of independent steady-state structural models of photosynthetic light reactions, the Calvin cycle and glycolytic reactions of the chloroplast, cytosolic glycolysis and the TCA cycle. The approach used here is to make no attempt to simplify the topology of the system, and to use the software ScrumPy described in Chapter 1 to define each of these models with a structure as complete as knowledge of the system allows. Once defined, the behaviour and properties of these models were investigated using structural analysis techniques such as ESs analysis and EM analysis.

An important aspect to be considered while constructing models of plant metabolism is the segregation of reactions and metabolites within specific subcellular compartments (Section 2.2). Considering the localisation is of no (or less) significance in case of small metabolic models representing reactions occurring in a single compartment. However, in large metabolic models spanning multiple compartments, the localisation of reactions and metabolites has to be carefully considered and represented as many of them can exist in multiple compartments. In the case of the modelling described in this thesis, reactions and metabolites in particular compartments were segregated by simply attaching a suffix containing the first three letters of the compartment to the original name. For example, '_lum', '_str', '_cyt', '_mit' and '_ims' were used to suffix components localised in lumen, stroma, cytosol, mitochondria and IMS, respectively.

Unless stated otherwise, models were reconstructed from the latest available biochemical literature sources. The stoichiometries of some of the reactions were obtained from the online pathway databases KEGG and AraCyc, and from previous models of plant metabolism constructed in our group [63, 181].

3.2 Model of photosynthetic light reactions

3.2.1 Model definition

A stoichiometric model of light reactions representing the photosynthetic electron transport chain was reconstructed manually. The model in ScrumPy format is available in Appendix A and is illustrated in Figure 3.1. The final version of the model contains 10 reactions and 15 metabolites.

The transfer of electrons through electron carriers leading to the formation of ATP and NADPH during photosynthetic light reactions forms the key structure of the network. Photo-activated oxidation of water and concurrent release of electrons is represented in the model as a property of PSII. Subsequent oxidation of PSII and the ensuing electron transport chain is represented by a series of coupled oxidation-reduction reactions. PSI is included in the model as a photo-activated redox reaction transferring electrons from PC to FD. The number of photons consumed during the PSI and PSII reactions was based on the evidence furnished in a recent review [112]. Although involved in the ETC, some electron carriers such as pheophytin and phylloquinone were not included in the model so as to reduce its size and complexity. The final electron acceptor NADP was linked to the model through the membrane-bound FD oxidoreductase reaction that converts $NADP^+$ to NADPH. Reverse electron flow in the ETC is represented in the model using a redox reaction ('Cyclic_lum') involving FD and Quinone A (QA).

An important feature of the model is that it represents the exchange of protons between two compartments — stroma and lumen. Two protons are transported into the lumen when an electron moves from water to FD. The movement of protons from the lumen to stroma was represented in the model by an ATPase reaction that translocates 14 protons to produce three ATP [112]. Two new reactions ADPSy and NADPHOx were included in the stromal fraction of the model to regenerate ADP and NADP (see Appendix A).

Photons, H_2O and O_2 were declared as external metabolites (Section 1.2.1) as they are in constant exchange with the external environment. Similarly, stromal protons were declared external as the model assumes the ETC on the thylakoid membrane as the boundary of the system. Apart from these, two additional metabolites 'ATPWork' and 'NADPHWork' were introduced and declared external to aid the investigation of the

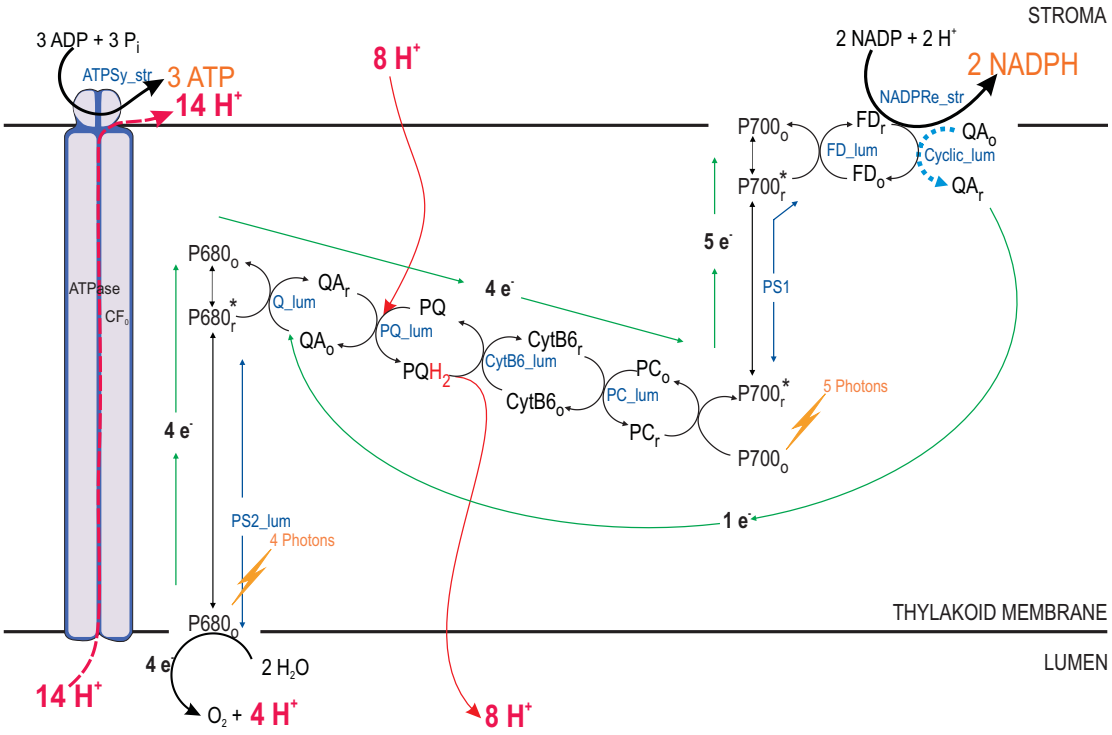


Figure 3.1 – Reaction schematic of the model of the photosynthetic electron transport chain. See List of Abbreviations for metabolite abbreviations and Appendix A for the stoichiometries of reactions. Red and green arrows indicate the direction of flow of protons and electrons, respectively. Reaction names are indicated with blue text. Metabolite and reaction name suffixes ('_lum' and '_str') and stromal sink reactions (ADPSy and NADPHOx) are not shown here for clarity, but were included in the model.

model with respect to the production and consumption of ATP and NADPH. The major purpose of using these dummy external metabolites was to directly derive net energy and reducing yields from the net stoichiometries of the EMs to be generated later.

3.2.2 Model analysis

No dead-end metabolites or dead reactions were revealed during the initial analysis of the model, and all reactions were found to be atomically balanced. Subsequent enzyme subsets analysis revealed four subsets containing more than one reaction and one involving only a single reaction, as shown in Table 3.1. While the reaction involved in reverse electron flow made a single subset of its own, reactions involved in the production of ATP and NADPH formed independent subsets with sink reactions that are responsible for the regeneration of ADP and NADP, respectively. The remaining subsets involving PS1 and PS2 (ESs 1 and 4, respectively) form the basic structure of the network. ES 1 grouped reactions that are involved in electron transfer from PSII to FD. ES 4, on the other hand, grouped reactions involved in the oxidation of water and the initiation of ETC.

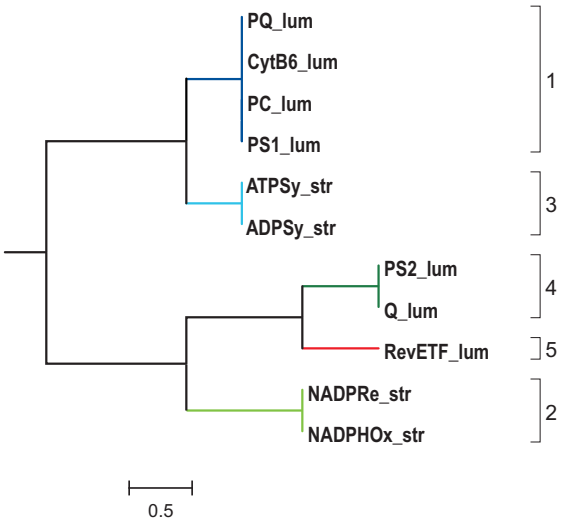


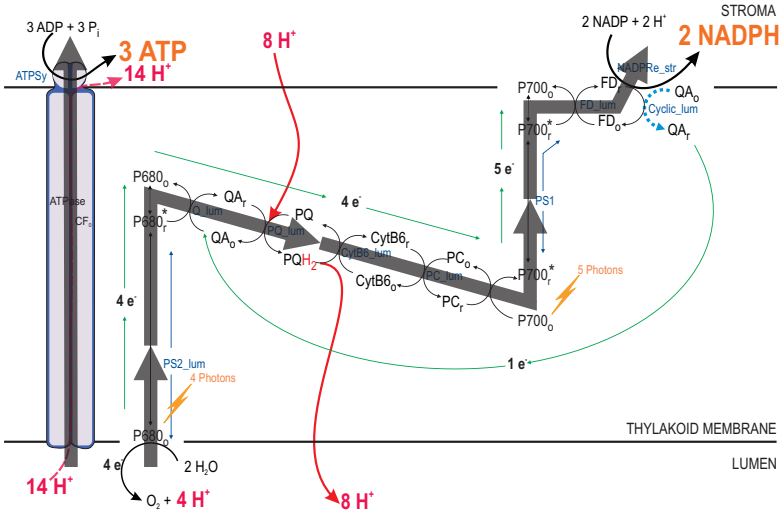
Figure 3.2 – Metabolic tree representing the model of the photosynthetic electron transport chain. See List of Abbreviations for metabolite abbreviations and Appendix A for the stoichiometries of reactions. The scale bar respresents a difference of $\theta_{xy}^{K} = 0.5$ rad.

A metabolic tree based on reaction correlation coefficients was constructed from the orthogonal null space of the stoichiometry matrix of the ETC model and is shown in Figure 3.2. Leaves of this tree represent reactions and nodes represent clusters that indicate the correlation between fluxes carried by reactions in the model. For example, the fluxes through reactions in Clusters 1 and 3 strongly correlate with each other as there is no apparent distance between them. Furthermore, note that the clusters on this tree correlate with the ESs shown in Table 3.1. Enzymes in same subset have a RCC of one, i.e. they perfectly correlate. The biological significance of the ESs and the metabolic tree will be discussed elsewhere (Section 3.2.3).

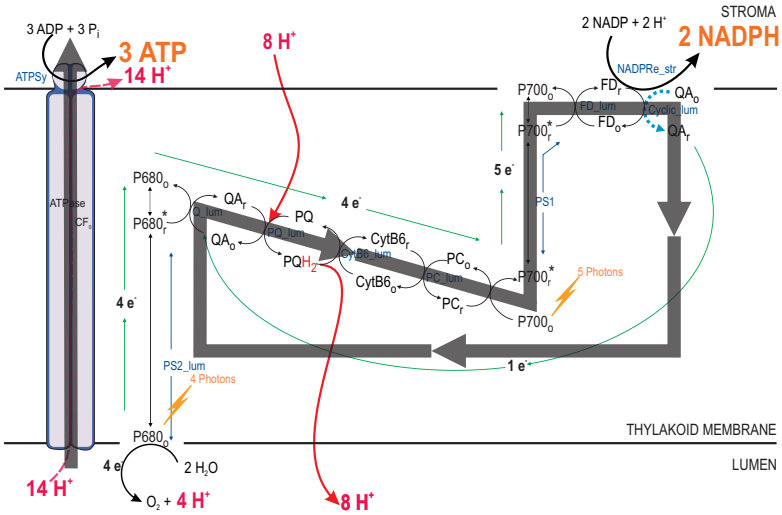
Further analysis of the model using the EMs algorithm generated two EMs, the overall stoichiometries of which are shown in Table 3.2. The structure of these EMs is overlaid on the network diagram from Figure 3.1 in Figure 3.3. The dummy external

Table 3.1 – Enzyme subsets in the model of the photosynthetic electron transport chain. See Figure 3.1 for a graphical representation of the reactions involved. Stoichiometries of the reactions are available in Appendix A.

Subset	Reactions	Function
1	PS1_lum CytB6_lum PQ_lum PC_lum	Redox reactions of the electron transport chain
2	NADPRe_str NADPHOx_str	Production of NADPH and regeneration of NADP
3	ADPSy_str ATPSy_str	ATP production and regeneration of ADP
4	PS2_lum Q_lum	Water oxidation and initiation of electron transport
5	Cyclic_lum	Reverse electron flow through FD and QA



(a)



(b)

Figure 3.3 – Elementary modes of the model of the photosynthetic electron transport chain. (a) EM representing non-cyclic electron flow in the ETC leading to the production of ATP and NADPH. Note that this EM is capable of generating proton gradient in the lumen by importing protons (red arrows) from the stroma and by splitting water. ATP is produced when these electrons are pumped from the lumen to stroma by ATPase. NADPH is formed when the electrons are transferred from FD to NADP. (b) EM representing cyclic electron flow in the ETC. This EM contributes to the net ATP yield by importing additional protons via PQ. Here, electrons are transferred from FD to QA.

Table 3.2 – Stoichiometries of EMs of the photosynthetic ETC model. ATPWork and NADPHWork are dummy external metabolites that directly represent the net energy and reducing yields of the EMs. '_lum' and '_str' indicate lumen and stromal localisation of metabolites, respectively. The external species Proton_str is omitted here for clarity, but was included in the analysis.

EM	Substrates	Products
1	1 H_2O_lum + 7 Photons	$\frac{1}{2}$ O_2+ $\frac{9}{7}$ ATPWork + 1 NADPHWork
2	2 H_2O_lum + 9 Photons	1 O_2+ $\frac{12}{7}$ ATPWork

metabolites ATPWork and NADPHWork were used to directly derive net energy and reducing yields from the net stoichiometry of the EMs. Note that the stoichiometries of these EMs correspond to the cyclic and non-cyclic photophosphorylation in chloroplasts (Section 2.2.1.1).

3.2.3 Discussion

ESs analysis of the model revealed groups of reactions that always operate together and thus share strictly coupled fluxes. Several interesting observations were made from the list of ESs shown in Table 3.1. ES 1 grouped all those redox reactions that are involved in transferring electrons from PSII to FD, suggesting that they operate in fixed flux proportions. This justifies the initial act of not including in the model some of the intermediate redox reactions such as pheophytin and phylloquinone that are involved in the photosynthetic ETC in plants. Including them would have served only to lengthen the list of reactions in ES 1, thereby making further interrogation of the model much more complicated. More importantly, ES 1 contains the redox reaction PQ that is responsible for the generation of the proton gradient in the lumen. This property of ES 1 makes it an important set of reactions responsible for photosynthesis. An equally important, although smaller, subset is the ES 4 that contains reactions involved in the oxidation of water and the initiation of the ETC. Oxidation of water liberates protons needed to make up the total number required for the production of ATP, and electrons required to excite the chlorophyll molecules in PSII. ESs 2 and 3, on the other hand, grouped reactions that are involved in the production and regeneration of NADPH and ATP, respectively. The fluxes through these ES are maintained by the amount of ATP and NADPH produced. Another interesting subset is ES 5 that contains the reaction responsible for reverse electron flow. In the photosynthetic ETC, reverse electron flow from FD to QA is responsible for cyclic photophosphorylation and generation of additional ATP.

A reaction correlation tree represents the correlation between the fluxes carried by reactions in the steady-state model. Each of the ESs in Table 3.1 was found to form a separate cluster on the reaction correlation tree shown in Figure 3.2. It is evident that Clusters 1 and 3 representing reactions involved in electron transport and ATP synthesis, are tightly correlated. This result conforms to the scenario observed in photosynthetic ETC in plant cells, where PMF generated by protons pumped into the lumen during electron transport via PQ results in the formation of ATP (Section 2.2.1.1). Clusters 4 and 5 in the metabolic tree are also highly correlated, the biological significance of which is debated [112, 182]. One argument is that the reactions in these two clusters are required to make up the protons needed for ATP synthesis.

The major objective behind EM analysis was to interrogate the behaviour and properties of the model. The overall stoichiometries of the EMs 1 and 2 as shown in

Table 3.2 correspond to the cyclic and non-cyclic photophosphorylation in chloroplasts, respectively. Several interesting observations were made by considering the reactions involved in each of these EMs (Figures 3.3 (a) and (b)). Reactions in EM 1 carry out oxidation of water in the presence of light to form protons and O_2 in the lumen. In addition to this they participate in the ensuing electron transport chain that mediates the import of protons from stroma to lumen. For each electron carried through the ETC reactions, two protons are pumped into the lumen. Hence, in one iteration of EM 1, a total of 12 protons (four from the oxidation of two molecules of water + eight from electron transport through PQ) are pumped into the lumen. ATP is produced by this EM when protons are pumped back into the stroma through the ATP synthase protein complex. The CF_0 unit of the ATPase complex in plant chloroplasts requires 14 protons for a complete rotation resulting in the formation of three molecules of ATP (Section 2.2.1.1). Since only 12 protons are available in the lumen the ratio of net ATP/NADPH calculated from the stoichiometry of EM 1 is 2.57/2. This ratio agrees with the observations made by Allen (2002) [112] in the most recent review on this subject. NADPH is produced when electrons are transferred from FD to NADP.

The EM representing cyclic photophosphorylation (EM 2), however, only contains reactions that are involved in the ETC and the reactions responsible for the reverse electron flow. In this EM, an electron that reaches FD through the ETC is transferred back to QA. Subsequent movement of this electron through the ETC mediates the import of two protons from the stroma to the lumen. These protons contribute to the overall proton requirement for the synthesis of ATP. Based on overall stoichiometry, EM 2 can produce 24/14 ATP from two molecules of H_2O. Although the real function of cyclic photophosphorylation in chloroplasts is still debated, it follows from the above observations that it mediates the import of protons into the lumen, thereby increasing the rate of non-cyclic photophosphorylation.

3.3 The model of the Calvin cycle

3.3.1 Model extension

The stoichiometric model of the Calvin cycle constructed by Poolman *et al.* (2003) [63] (reviewed in Section 2.3) was modified to fit the requirements of this study by removing reactions involved in dark metabolism (OPPP) and by including additional reactions that are active in the presence of light. To begin with, glycolytic reactions in the chloroplast that mediate the sequential conversion of PGA to MAL and OAA were introduced into the model. This was followed by the addition of transport reactions responsible for the export of intermediates of the new reactions (MAL, OAA and PEP) and G6P into the cytosol (see Section 2.2.4.4 for a detailed description of these

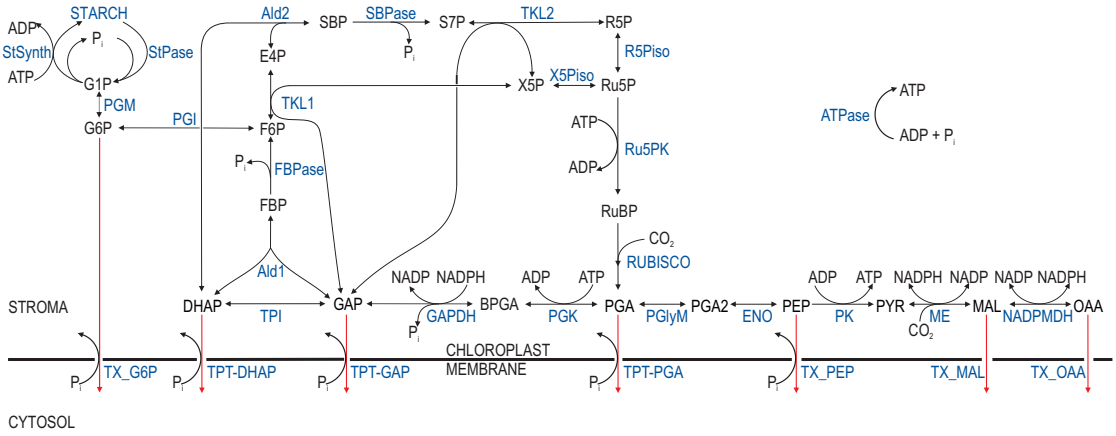


Figure 3.4 – Reaction schematic of the model of the Calvin cycle containing stromal 'glycolytic' reactions. See List of Abbreviations for metabolite abbreviations and Appendix B for the stoichiometries of reactions. Red arrows indicate transport reactions exporting intermediates of chloroplast metabolism to the cytosol. Reaction names are indicated with blue text. Metabolite and reaction name suffixes ('_str' and '_cyt') are not shown here for clarity, but were included in the model.

transport reactions). Export of stromal G6P and PEP were each coupled with the import of cytosolic P_i. To simplify future analysis and interrogation of the model the MAL/OAA shuttle was included as two independent reactions that export MAL and OAA to the cytosol. Light reactions were included as part of the original model in the form of a sink reaction producing ATP. NADPH and NADP were considered as external metabolites. The resulting model represented carbon fixation via CO_2 assimilation to produce starch, triose-phosphate intermediates of the Calvin cycle and intermediates of stromal glycolytic reactions, and subsequent export of some of these intermediates into the cytosol.

Apart from NADPH and NADP, CO_2, starch and cytosolic GAP, DHAP, PGA and P_i constituted the external metabolites in the original model. In addition to this, four new external metabolites were introduced into the extended model to represent the cytosolic versions of G6P, PEP, MAL and OAA exported from the chloroplast. The final version of the model, containing 30 reactions and 23 metabolites, is furnished in ScrumPy '.spy' format in Appendix B and is illustrated in Figure 3.4.

3.3.2 Model analysis

ESs analysis of the extended model revealed six subsets containing more than one reaction and 12 involving only a single reaction. A list of the reactions constituting the major ESs and their respective functions is shown in Table 3.3. The largest subset, ES 2, was composed of eight reactions and its biological significance will be discussed elsewhere (Section 3.3.3). The remaining ESs were found to be trivial and will not be discussed further.

A metabolic tree based on reaction correlation coefficients was constructed from the orthogonal null space of the stoichiometry matrix of the extended Calvin cycle model

Table 3.3 – Enzyme subsets of the extended Calvin cycle model containing more than two reactions. See Figure 3.4 for a graphical representation of the reactions involved. Stoichiometries of the reactions are available in Appendix B. '↔' and '→' indicate the reversible and irreversible conversion of one set of metabolites to another by reactions in that particular subset, respectively. Other less significant subsets that are composed of only a single reaction are omitted.

Subset	Reactions	Function
1	ME_str PK_str	PEP → MAL
2	Ald2_str Rubisco_str SBPase_str TKL2_str Ru5PK_str R5Piso_str TKL1_str X5Piso_str	Regenerative phase of the Calvin cycle GAP → RuBP
3	Ald1_str FBPase_str	F6P → triose-phosphates DHAP and GAP
4	NADPMDH_str TX_OAA_str	MAL ↔ OAA and the export of OAA from stroma to cytosol
5	GAPDH_str PGK_str	GAP ↔ PGA
6	Eno_str PGlyM_str	PGA ↔ PEP

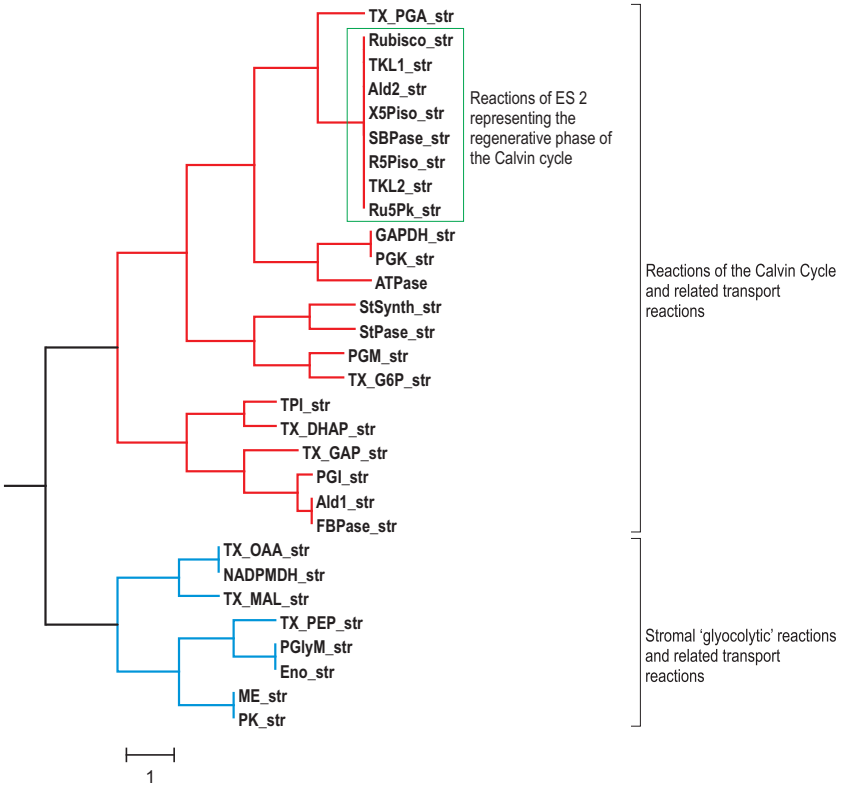


Figure 3.5 – Metabolic tree representing the extended Calvin cycle model. See List of Abbreviations for metabolite abbreviations and Appendix B for the stoichiometries of reactions. The green rectangle highlights the ES in Table 3.3. The scale bar respresents a difference of $\theta^K_{xy} = 1$ rad.

Table 3.4 – Overall stoichiometries of the eight new EMs of the extended Calvin cycle model. Stoichiometries of the EMs in the original model are shown in Table 2.2. '_str' and '_cyt' indicate stromal and cytosolic localisation of metabolites, respectively. External species NADP_str, NADPH_str and P_i_cyt are omitted here for clarity, but were included in the analysis.

EM	Substrates	Products
1	6 CO_2	G6P_cyt
2	3 CO_2	PEP_cyt
3	4 CO_2	MAL_cyt
4	4 CO_2	OAA_cyt
5	3 CO_2 + Starch_str	3 PEP_cyt
6	6 CO_2 + Starch_str	3 MAL_cyt
7	6 CO_2 + Starch_str	3 OAA_cyt
8	Starch_str	G6P_cyt

(Figure 3.5). Two distinct clusters, one representing reactions of the Calvin cycle (red) and the other representing glycolytic reactions (blue), were observed in this tree. The node representing the largest ES is highlighted with a green box.

EM analysis of the extended model generated 16 EMs, half of which corresponded to the complete set of EMs obtained from the original model by Poolman *et al.* (2003) [63]. Overall stoichiometries of the EMs in the original model are tabulated in Table 2.2 and those of the additional new EMs obtained from the extended model are shown in Table 3.4. The biological significance of these EMs will be discussed in the next section.

3.3.3 Discussion

The major objective of ES analysis of the extended Calvin cycle model was to identify reactions that operate together in fixed flux proportions. It was found that the model is composed mainly of smaller ESs constituting either one or two reactions. However, the largest subset, ES 2, was composed of eight reactions that form the regenerative limb of the Calvin cycle. These reactions regenerate RuBP from GAP thereby ensuring the continuous assimilation of CO_2 into carbohydrates (Section 2.2.1.2). ES 2 can carry flux in only one direction, i.e. from GAP to RuBP, since several key reactions in this subset, such as Rubisco, Ru5PK and SBPase, are irreversible.

The reasons for performing EM analysis on the extended model were twofold: to interrogate the behaviour and properties of the model and to identify potential routes whereby carbon assimilated via the Calvin cycle is exported to the cytosol. A total of 16 EMs were generated from the extended model. Overall stoichiometries and participating reactions of eight of these EMs were identical to those from the original model of the Calvin cycle (Table 2.2). These EMs were classified as: (a) three EMs producing one each of PGA, DHAP and GAP from three CO_2; (b) three EMs producing three each of

PGA, DHAP and GAP from three CO_2 and G6P from starch; (c) one EM producing starch from CO_2; (d) a futile cycle synthesising and degrading starch. Using these EMs it was shown that starch degradation can serve to support the Calvin cycle in the presence of light by producing triose-phosphate intermediates that are exported to the cytosol [63]. Other interesting biological properties of these EMs were reviewed in Section 2.3. From the point of interest of this thesis, however, these EMs are considered as the various routes through which carbon fixed into Calvin cycle intermediates GAP, DHAP and PGA using energy and redox equivalents from light reactions are exported to the cytosol.

The remaining eight novel EMs of the extended model shown in Table 3.4 can be classified as: (a) one EM (EM 1) involved in the synthesis and export of G6P from CO_2; (b) three EMs (EMs 2-4) producing PEP, MAL and OAA from CO_2 and exporting them into cytosol; (c) three EMs (EMs 5-7) producing PEP, MAL and OAA from CO_2 and intermediates of starch degradation; (d) one EM (EM 8) involved in the synthesis and export of G6P liberated during starch degradation. A general characteristic of these EMs is that they represent the export of carbon to the cytosol. EMs 1-4 fix carbon via CO_2 assimilation into G6P, PEP, MAL and OAA and then export these intermediates to cytosol. EMs 5-8, on the other hand, use both CO_2 assimilation and starch degradation to fix carbon into the intermediates before exporting them to the cytosol. Another general feature of these EMs is that all of them, except EM 8, contain reactions constituting ES 2. This ES 2 backbone was also found to exist in the EMs of the original model, other than in the futile cycle. This observation suggests that the regenerative limb of the Calvin cycle is the most important set of reactions in Calvin cycle. Furthermore, it implies a strong coupling of ES 2 with the production of intermediates of the Calvin cycle.

3.4 Model of cytosolic glycolytic reactions

3.4.1 Model construction

A stoichiometric model of plant central carbon metabolism containing reactions involved in glucose catabolism and sucrose synthesis in the cytosol was constructed. The purpose of constructing this model was to study the distribution of carbon flux between the intermediates of chloroplast metabolism exported to the cytosol and the intermediates of gluconeogenesis and glycolysis such as sucrose, PYR, MAL and OAA. The final version of the model containing 26 reactions and 17 metabolites is available in ScrumPy format in Appendix C. An illustration of the model is shown in Figure 3.6.

Anabolic and catabolic reactions of the cytosol leading to the formation of sucrose and PYR, OAA and MAL, respectively, from intermediates of the Calvin cycle and

Table 3.5 – Enzyme subsets of the model of cytosolic glycolytic reactions containing more than two reactions. See Figure 3.6 for a graphical representation of the reactions involved. Stoichiometries of the reactions are available in Appendix C. '↔' indicates the reversible conversion of one set of metabolites to another by reactions in that particular subset.

Subset	Reactions	Function
1	TX_MAL_cyt NADMDH_cyt	Synthesis and exchange of MAL
2	PGM_cyt NDPK_cyt SuSyn_cyt UGPase_cyt Suc_EX	Reactions involved in sucrose synthesis and export
3	PGlyM_cyt Eno_cyt	PGA ↔ PEP
4	PGK_cyt GAPDHP_cy	GAP ↔ PGA
5	TX_PYR_cyt PK_cyt	Synthesis and exchange of PYR

stromal glycolytic reactions, form the key structure of the network. Synthesis and degradation of sucrose was represented as a single reaction catalysed by the enzyme sucrose synthase. Activities of other sucrose degrading enzymes (Inv and HK) were not included in the model so as to simplify future analysis and interrogation. Sucrose metabolism was linked to PYR synthesis by means of standard glycolytic reactions that convert F6P to PYR. This branch of glycolysis was then attached to MAL/OAA metabolism by means of the CO_2-binding PEPC reaction. The model contains four irreversible reactions catalysed by enzymes PFK, GAPDH, PK and PEPC.

Import of the intermediates of chloroplast metabolism into the cytosol was represented in the model using specific transport reactions that mediate P_i-dependent antiport of G6P, DHAP, GAP, PGA and PEP. MAL and OAA exchange with stroma was implemented using two independent transport reactions. Similarly, a simple transport reaction was introduced into the model to represent the diffusion of PYR across the chloroplast membrane. Apart from the stromal metabolites associated with the above transport reactions, sucrose, protons, CO_2 and stromal P_i were declared as external metabolites. Additionally, ATP, ADP, NADH and NAD were declared external as the energy and reducing equivalents produced during chloroplast and/or mitochondrial metabolism were not available to the model. See Section 2.2.4.2 for a brief review on the mechanisms that mediate the import of ATP and NADPH into the cytosol. Compartmentation of metabolites and reactions in the cytosol was implemented by attaching the suffix '_cyt' to their names.

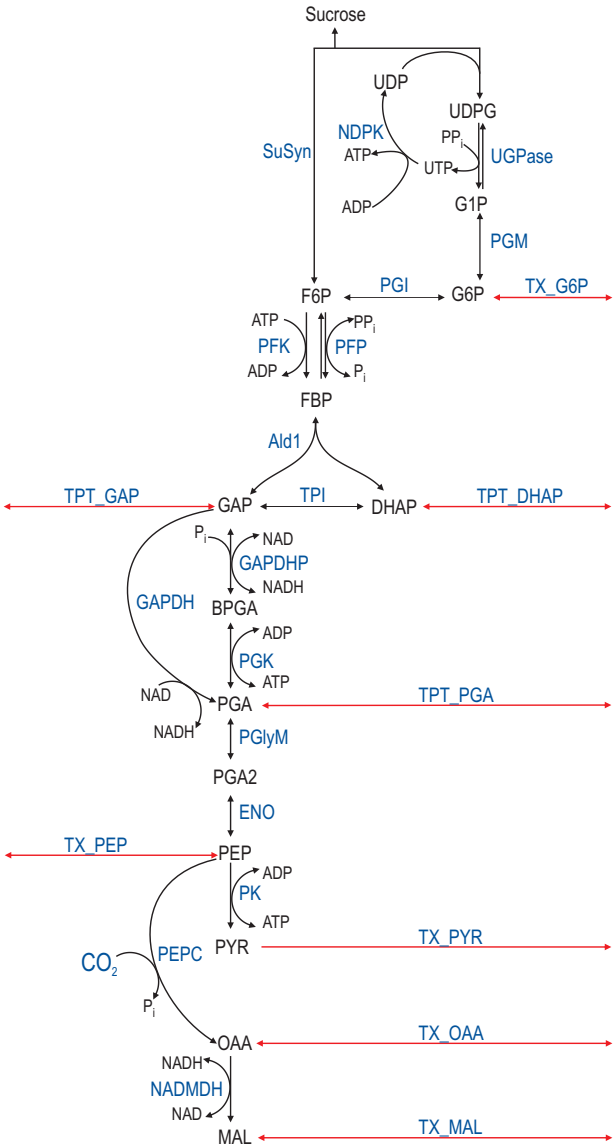


Figure 3.6 – Reaction schematic of the model of cytosolic glycolytic reactions. See List of Abbreviations for metabolite abbreviations and Appendix C for the stoichiometries of reactions. Red arrows indicate transport reactions exchanging intermediates of chloroplast metabolism to the cytosol. Reaction names are indicated with blue text. Metabolite and reaction name suffixes ('_str' and '_cyt') are not shown here for clarity, but were included in the model.

3.4.2 Model analysis

No dead-end metabolites or dead reactions were revealed during the initial analysis of the null space of the stoichiometry matrix representing the model. Subsequent enzyme subsets analysis revealed 14 subsets involving only a single reaction and five containing more than one reaction. A list of reactions constituting the major ESs and their respective functions is shown in Table 3.5. Smaller subsets are not listed in the table for clarity. The largest ES, ES 2, contained five reactions that act in fixed flux proportions to synthesise sucrose from G6P. NADMDH (ES 1) and PK (ES 5) formed independent subsets with transport reactions that mediate the exchange of their products

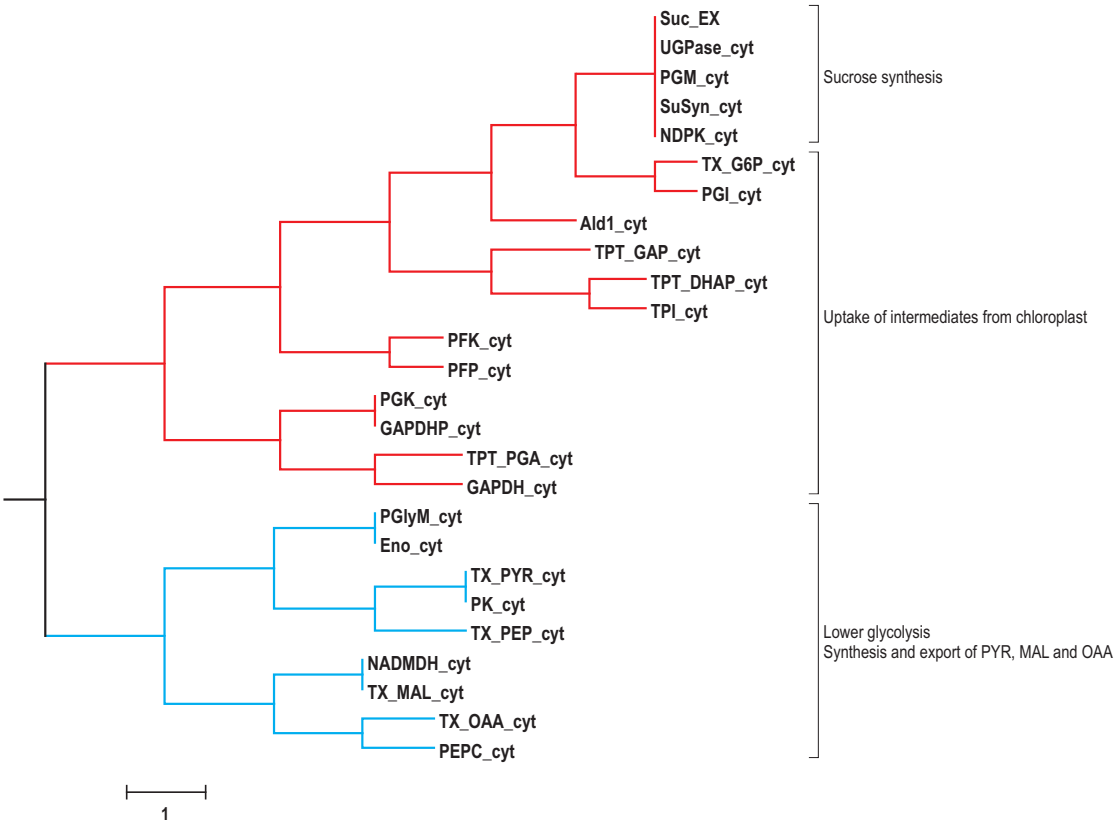


Figure 3.7 – Metabolic tree representing the model of glycolytic reactions in the cytosol. See List of Abbreviations for metabolite abbreviations and Appendix C for the stoichiometries of reactions. The scale bar respresents a difference of $\theta^K_{xy} = 1$ rad.

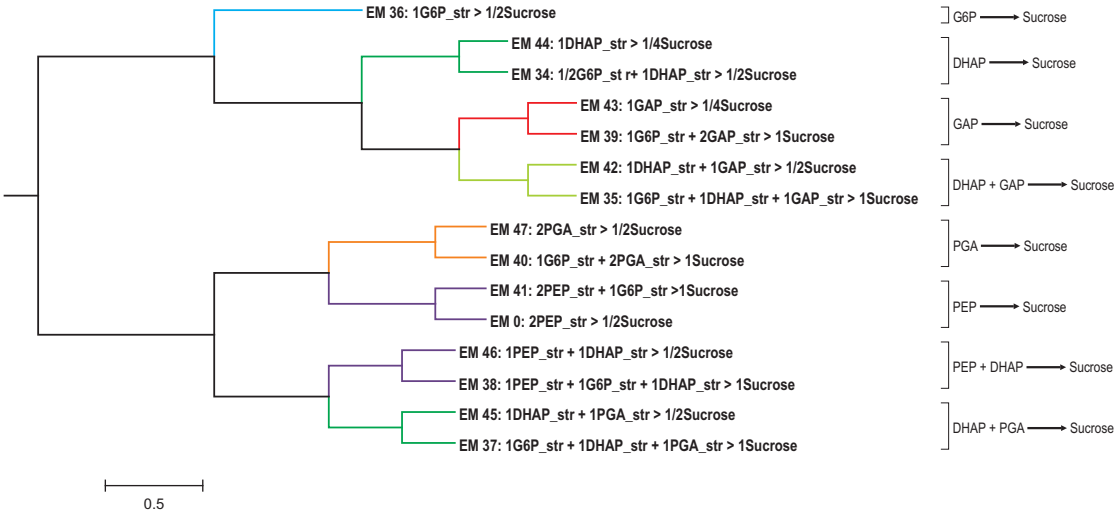


Figure 3.8 – Dendrogram showing overall stoichiometries of the EMs of the model of cytosolic glycolytic reactions that produce sucrose from chloroplast intermediates, clustered by angle based on their net external metabolite usage (Section 1.4.2.3). Some external metabolites were omitted here for clarity, but were included in the analysis. Branch colouration indicates differing net carbon stoichiometries. '_str' indicates the localisation of the external metabolites in the stroma. The scale bar respresents a difference of $\theta^K_{xy} = 0.5$rad.

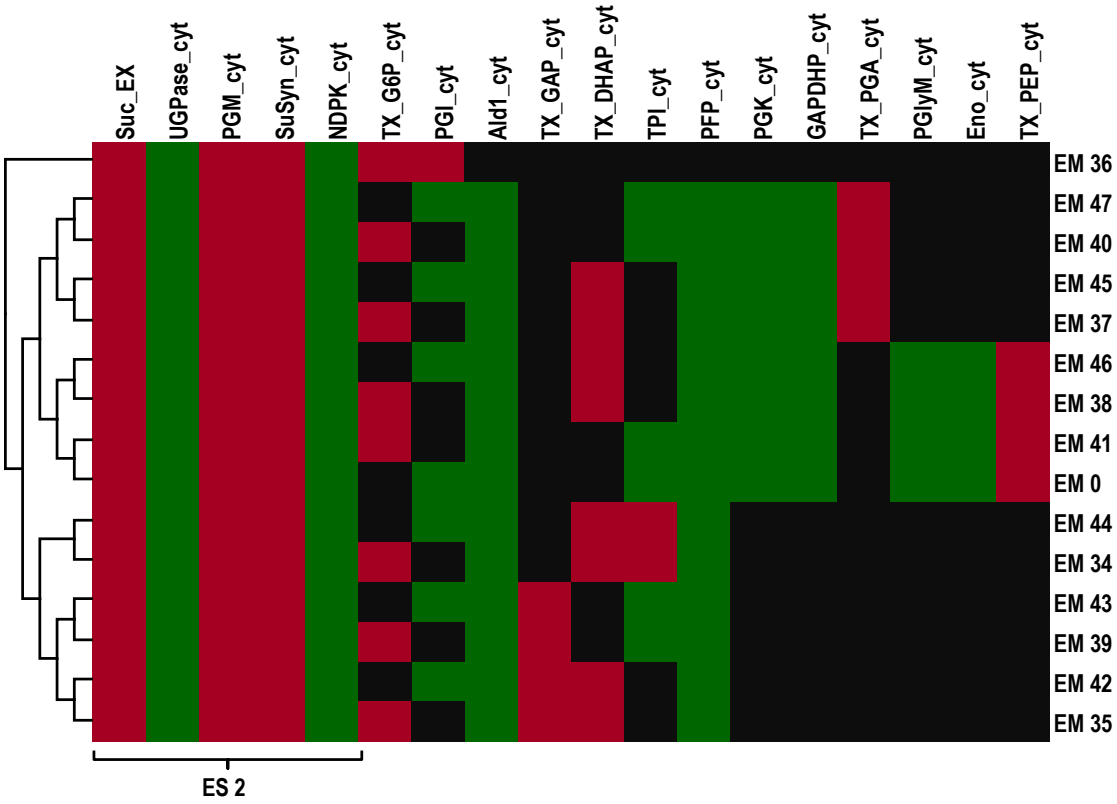
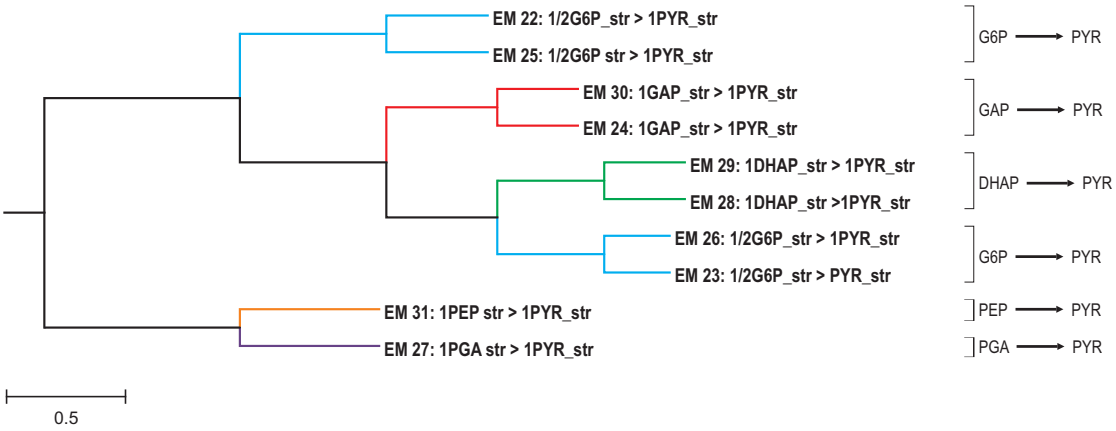


Figure 3.9 – Heatmap representing reactions participating in the EMs of the model of cytosolic glycolytic reactions that produce sucrose from chloroplast intermediates. Rows were clustered by angle based on their reaction usage. '_cyt' indicates localisation of reactions in the cytosol. The heatmap has been coloured according to the stoichiometric coefficient of a reaction in an EM, i.e. $= 0$ (black), < 0 (green) and > 0 (red).

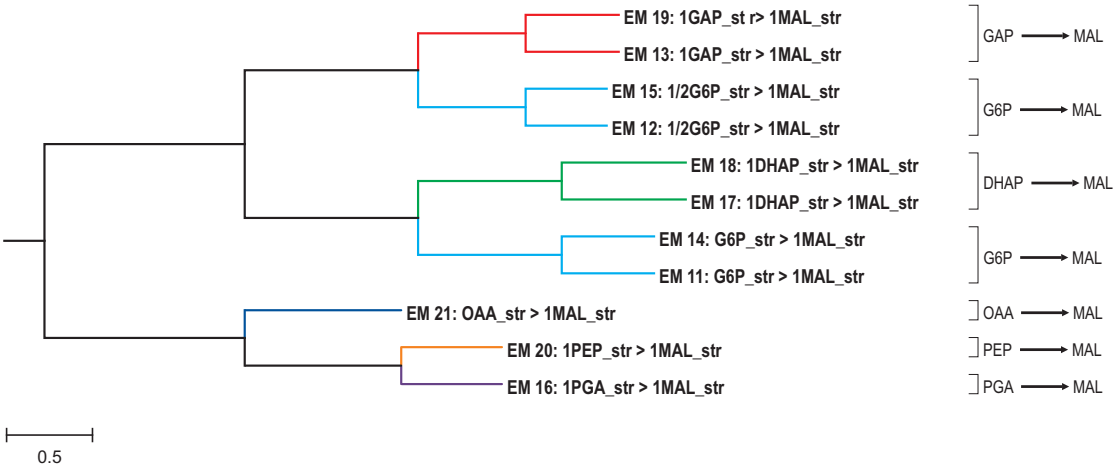
with the stroma. Reactions constituting ES 3 and ES 4 convert PGA to PEP and GAP to PGA, respectively. Note that these two ESs are identical to those obtained from the model of stromal carbon metabolism (Table 3.3).

A metabolic tree representing correlation between fluxes carried by reactions in the model was constructed from the stoichiometry of the model (Figure 3.7). Two distinct clusters were observed. The larger cluster (red) represented reactions of the upper glycolysis that perform two different, but related, tasks - importing intermediates of chloroplast metabolism into the cytosol and converting these intermediates into sucrose. The second cluster (blue) was composed of reactions of lower glycolysis that are involved in the synthesis and subsequent export of PYR, MAL and OAA.

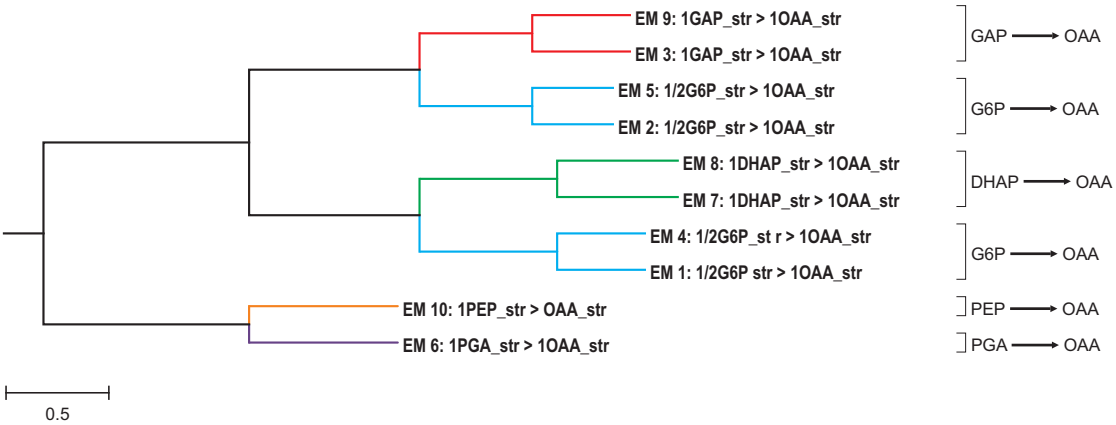
The major objective of performing EM analysis on this model was to ascertain the existence of carbon flux from intermediates of chloroplast metabolism to sucrose, and PYR, MAL and OAA. A total of 48 EMs representing various routes mediating such carbon flux were generated from the model. From the entire set of EMs, four independent subsets containing modes responsible for the production of sucrose, PYR, MAL and OAA were extracted. EMs in each of these subsets were hierarchically clustered based on the net usage of external metabolites. The reason for doing this was to group together EMs that share similar functional properties.



(a)



(b)



(c)

Figure 3.10 – Dendrogram showing overall stoichiometries of EMs of the model of cytosolic glycolytic reactions that produce PYR (a), MAL (b) and OAA (c) from chloroplast intermediates clustered by angle based on their net external metabolite usage. Some external metabolites were omitted here for clarity, but were included in the analysis. Branch colouration indicates differing net carbon stoichiometries. '_str' indicates localisation of metabolites in the stroma. Scale bars respresents a difference of $\theta^{K}_{xy} = 0.5$rad.

A dendrogram illustrating the arrangement of clusters produced by hierarchical clustering of the stoichiometries of EMs producing sucrose is shown in Figure 3.8. A heatmap representing reactions participating in the EMs was constructed from the EMs reaction matrix and the rows were hierarchically clustered (Figure 3.9). Columns of the heatmap were then sorted with respect to the order of leaves on the reaction correlation tree shown in Figure 3.7. The presence of EMs converting G6P, DHAP, GAP, PGA and PEP to sucrose is evident from the overall stoichiometries of EMs shown in Figure 3.8 and also from the reaction participation heatmap. Meanwhile, dendrograms representing the overall stoichiometries of EMs mediating carbon flux from chloroplast intermediates to PYR, MAL and OAA are shown in Figures 3.10 (a), (b) and (c), respectively.

3.4.3 Discussion

Results from the ESs analysis, shown in Table 3.5, reveal some very interesting properties of the model. The major ES, ES 2, containing reactions involved in sucrose synthesis and export, represents the only means by which sucrose can be synthesised by the model. Therefore, synthesis of sucrose from any intermediate of chloroplast metabolism must involve this subset. This is evident from the fact that all EMs in the system that produce sucrose use this ES, as shown in the reaction participation heatmap in Figure 3.9. Furthermore, in the reaction correlation tree representing the model (Figure 3.7), it can be seen that flux through ES 2 correlates with those through PGI and the G6P transporter. These observations suggest that for any G6P entering the system, there is a higher chance of it to be converted to sucrose. Further analysis using more sophisticated approaches such as kinetic modelling or FBA must be performed before considering this any further. EM analysis of the model, however, generated an EM that could produce sucrose directly from G6P using ES 2 and PGI (EM 36 in Figures 3.8 and 3.9). Another interesting aspect here is the importance of enzymes aldolase (Ald1) and PFP that convert GAP and DHAP to FBP and FBP to F6P, respectively, for sucrose synthesis. As can be easily seen in the heatmap, they are involved in all but one EM that produces sucrose from chloroplast intermediates, thus signifying their importance in sucrose metabolism.

Another important aspect that was investigated using EM analysis was the fate of intermediates of chloroplast metabolism imported into the cytosol. On this basis, the entire set of EMs obtained from the model may be divided between as those producing sucrose via the gluconeogenesis pathway and those producing PYR, MAL and OAA via the glycolysis pathway. The former is exemplified by the overall stoichiometries of EMs shown in Figure 3.8, indicating carbon flux from chloroplast intermediates G6P, GAP, DHAP, PGA and PEP (and a combination of some of these intermediates) to

sucrose. Carbon flux from PYR, MAL and OAA to sucrose was not identified in the EMs because of the presence of two irreversible reactions PK and PEPC (Figure 3.6). On the other hand, overall stoichiometries of EMs producing PYR, MAL and OAA from the aforementioned chloroplast intermediates represented in dendrograms 3.10 (a), (b) and (c), respectively, indicate carbon flux via the glycolytic pathway.

3.5 Model of the TCA cycle and oxidative phosphorylation

3.5.1 Model definition

A structural model of mitochondrial metabolism composed of the reactions of the TCA cycle and the associated electron transport chain was constructed manually. Synthesis of redox equivalents during the sequential conversion of PYR to OAA via reactions of the TCA cycle and concurrent ATP synthesis, mediated by an electron transport chain initiated by the oxidation of these redox equivalents, forms the key structure of the model. The major objective of constructing this model was to investigate the various routes through which PYR is converted to NADH and ATP, and ascertain the existence of carbon flux. The final version of the model containing 17 reactions and 21 metabolites is available in ScrumPy '.spy' format in Appendix D and is illustrated in Figure 3.11.

The model represents interaction between three compartments — cytosol, mitochondria and the IMS. Import of PYR from the first compartment was represented in the model using a transport reaction. Subsequent sequential reactions converting PYR to OAA epitomised the TCA cycle division of the model. Of these, four reactions catalysed by enzymes PDH, IDH, AKGDH and NADMDH were accompanied by the synthesis of NADH. CoA-SH, CO_2 and H_2O associated with some reactions were not considered in the model definition in order to simplify future analysis and interrogation. The second division of the model is comprised of the reactions of the mitochondrial ETC. Oxidation of NADH and the resulting initiation of the ETC were represented in the model using the Complex I reaction. The Complex II reaction, on the other hand, initiated the ETC during the oxidation of succinate to fumarate. Export of protons into the third compartment, the IMS, was represented in the model by the reactions of Complex I, III and IV, and the resulting ATP synthesis (when protons return to the matrix) was represented by Complex V.

The MAL/OAA shuttle on the mitochondrial membrane was included in the model as two independent reactions pumping MAL and OAA to the cytosol. The mitochondrial AKG transporter was not included in the analysis as cytosolic nitrogen metabolism is outside the scope of this thesis. Apart from PYR, MAL and OAA, two additional external metabolites 'NADHWork' and 'ATPWork' were introduced into the model to

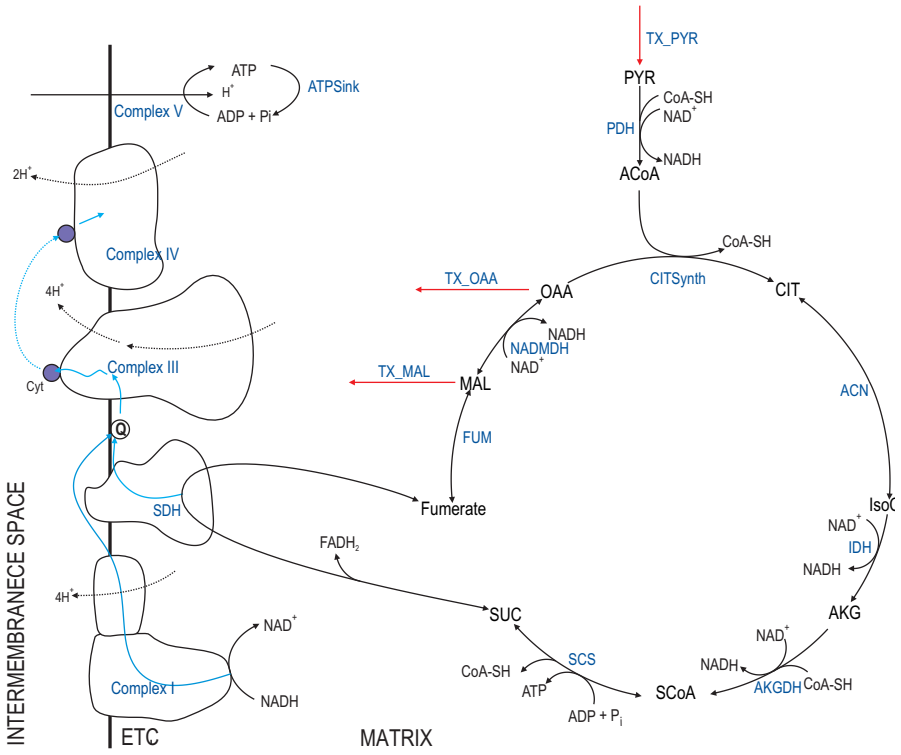


Figure 3.11 – Reaction schematic of the model of mitochondrial metabolism. See List of Abbreviations for metabolite abbreviations and Appendix D for the stoichiometries of reactions. Red arrows indicate transport reactions. Reaction names are indicated with blue text. Metabolite and reaction name suffixes ('_mit' and '_cyt') are not shown here for clarity, but were included in the model.

investigate the production and consumption of ATP and NADH directly from the net stoichiometries of EMs. Protons in the intermembrane space were also declared external.

3.5.2 Model analysis

Enzyme subsets analysis of the model revealed three subsets containing more than one reaction and four involving only a single reaction. A list of reactions constituting the major ESs and their respective functions is shown in Table 3.6. The largest subset, ES 1, constituted nine reactions of the TCA cycle that convert PYR in the mitochondrial matrix to MAL. Subsequent conversion of MAL to OAA, however, also depends on transport reactions of the MAL/OAA shuttle that form ES 3. The remaining subset, ES 2, grouped together reactions of the mitochondrial ETC that mediate electron transport through cytochrome and resulting proton import into the IMS.

A metabolic tree based on reaction correlation coefficients was constructed to study the correlation between fluxes carried by reactions in the model and is shown in Figure 3.12. Two distinct clusters, one representing reactions of the TCA cycle and the other representing reactions of the mitochondrial ETC, were observed in the metabolic tree. The former cluster was further divided into two separate clusters representing reactions in ES 1 and ES 3, respectively. A cluster representing the mitochondrial

Table 3.6 – Enzyme subsets of the model of mitochondrial metabolism containing more than two reactions. See Figure 3.11 for a graphical representation of the reactions involved. Stoichiometries of the reactions are available in Appendix D. Other less significant subsets that are composed of only a single reaction are omitted.

Subset	Reactions	Function
1	FUM_mit CITSynth_mit SCS_mit SDH_mit ACN_mit AKGDH_mit TX_PYR_mit IDH_mit PDH_mit	Reactions of the TCA Cycle converting PYR to MAL
2	Complex_III Comples_IV	Electron transport through cytochrome and proton export
3	TX_MAL_mit TX_OAA_mit	MAL and OAA exchange with cytosol

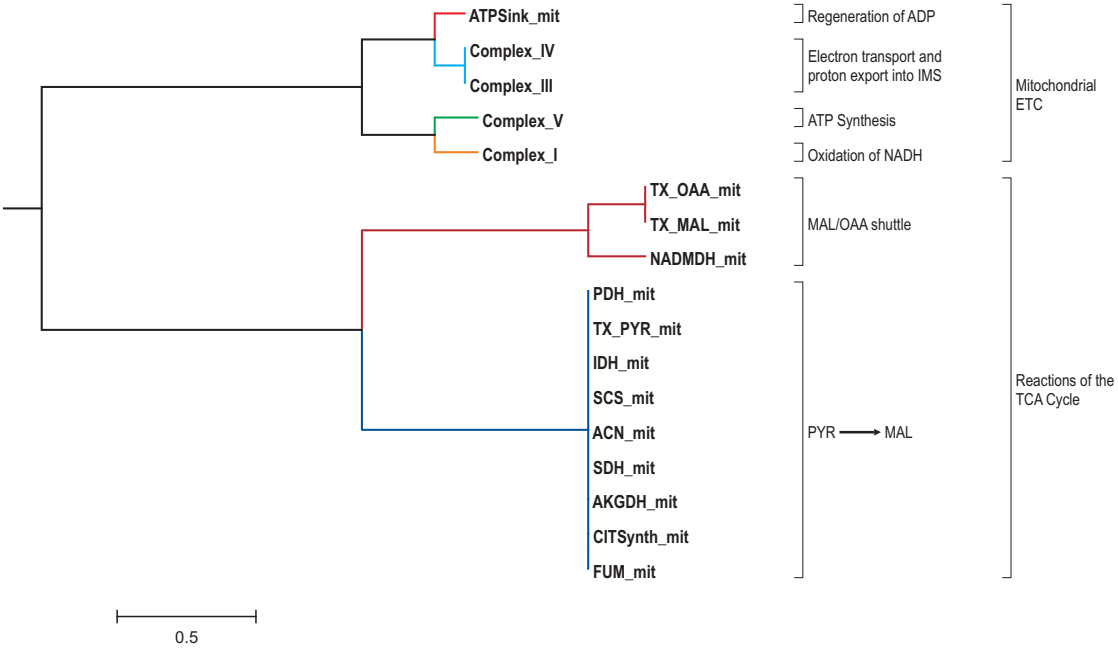


Figure 3.12 – Metabolic tree representing the model of mitochondrial metabolism. See List of Abbreviations for metabolite abbreviations and Appendix D for the stoichiometries of reactions. The scale bar respresents a difference of $\theta^K_{xy} = 1$ rad.

ETC, on the other hand, had four separate sub-clusters, the largest of which contained the reactions Complex_III and Complex_IV that mediate electron transfer to the final electron acceptor. Complex_I and Complex_V formed related but distinct clusters on the tree.

EM analysis of the model revealed four EMs, the overall stoichiometries of which are shown in Table 3.7. A heatmap representing reactions participating in each of these EMs was constructed from the EMs reaction matrix by hierarchically clustering its rows (Figure 3.13). EM 1 involves the synthesis of ATP and redox equivalents from PYR imported into the mitochondrial matrix from cytosol. This EM represents the oxidation

Table 3.7 – Overall stoichiometries of the EMs of the model of mitochondrial metabolism. 'ATPWork' and 'NADHWork' represent ATP formed and NAD^+ regenerated, respectively. External metabolite Proton_ims (proton localised in IMS) is omitted here for clarity, but were included in the analysis. '_cyt' and '_mit' indicate cytosolic and mitochondrial localisation of metabolites, respectively.

EM	Substrates	Products
1	1 PYR_cyt	4 NADHWork_mit + $\frac{11}{2}$ ATPWork_mit
2	2 PYR_cyt + 2 OAA_cyt	MAL_cyt + 6 NADHWork_mit + 9 ATPWork_mit
3	2 PYR_cyt + 8 OAA_cyt	8 MAL_cyt + 3 ATPWork_mit
4	2 MAL_cyt	2 OAA_cyt + 2 NADHWork_mit + 2 ATPWork_mit

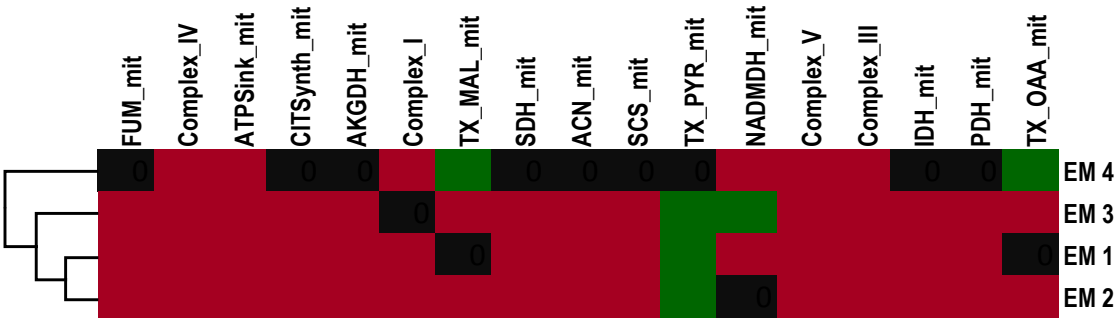


Figure 3.13 – Heatmap representing reactions participating in the EMs of the model of mitochondrial metabolism. Rows were clustered by angle based on their reaction usage. '_mit' indicates localisation of metabolites in the cytosol. The heatmap has been coloured according to the stoichiometric coefficient of a reaction in an EM, i.e. $= 0$ (black), < 0 (green) and > 0 (red).

of NADH produced by reactions of the TCA cycle and the production of ATP mediated by the resulting ETC. EMs 2 and 3 produce ATP from PYR and OAA imported from the cytosol. The difference between these two modes is that EM 3 does not involve the production of reducing equivalents. The remaining EM, EM 4, represents ETC mediated ATP generation initiated by the oxidation of NADH produced via the MAL/OAA shuttle without the need for any carbon flux.

3.5.3 Discussion

ESs analysis of the model revealed some very important properties of the system. The largest ES (Table 3.6), ES 1, represents the major path of carbon flux in the model leading up to the synthesis of MAL from cytosolic PYR imported into the mitochondria. It follows that the reactions of this subset produce most of the reducing equivalents generated by the model. Additionally, the presence of SDH in this subset means that any flux through it will lead to the formation of ATP via the ETC. Similar inferences can be drawn from the heatmap representing the reactions participating in the EMs of the model (Figure 3.13). Here, EM 3 uses reactions of ES 1 and the ETC initiated by SDH to produce ATP from PYR. EMs 1 and 2, however, uses the set of reactions used by EM 3 along with Complex_I to produce ATP and regenerate NAD^+.

The other major ES, ES 3, contains transport reactions responsible for the exchange MAL and OAA. From the reaction correlation tree shown in Figure 3.12, it is evident that the fluxes through the reactions in ES 3 correlate very strongly with that through NADMDH reaction. These three reactions together form the basic structure of MAL/OAA shuttle that can indirectly import reducing equivalents into the mitochondrial matrix. Oxidation of NADH by Complex_I initiates the ETC that lead to the formation of ATP. EM 4 in Figure 3.13 represents the reactions that use this shuttle mechanism to produce ATP and NAD^+ in the mitochondrial matrix.

The primary reasons for performing EM analysis on this model were twofold: to investigate the carbon flux through reactions of the TCA cycle and to investigate the major routes of ATP synthesis. From the overall stoichiometries of the EMs shown in Table 3.7 and the reaction participation heatmap shown in Figure 3.13 , it is evident that EMs 1, 2 and 3 require oxidation of PYR to initiate carbon flux through the reactions of the TCA cycle. Along with ATP and the reducing equivalents, MAL and OAA are produced as part of this carbon flux. The major routes involved in ATP synthesis and the regeneration of NAD^+, however, were discussed in the previous paragraphs.

CHAPTER 4

Integrated models of plant metabolism

4.1 Introduction

The properties and behaviours of independent steady-state structural models of the light reactions, the Calvin cycle, glycolysis and the TCA cycle, representing metabolic pathways in the chloroplast, cytosol and mitochondria, were described in the previous chapter. The purpose of this chapter, however, is to investigate the characteristics of the interaction between these compartments using structural model analysis techniques.

From the review in Chapter 2, it must be noted that, unlike the mitochondrial membrane, the inner chloroplast membrane does not have dedicated ATP/ADP transporters that can export ATP to the cytosol. Nevertheless, ATP and NADPH formed in the stroma during the light reactions are exported to the cytosol for the synthesis of sucrose and fatty acids, and to drive metabolism in other organelles. To date, two shuttle mechanisms — triose-phosphate/PGA and MAL/OAA — have been identified that mediate the transfer of ATP and reducing equivalents from chloroplast to cytosol (see Section 2.2.4.2 for a detailed explanation). One of the original goals motivating this structural investigation of the interaction between compartments was to identify other potential routes or shuttle mechanisms through which ATP and NADPH may be exported from chloroplast to cytosol without any net carbon flux.

The role of mitochondrial metabolism in protecting plants from photoinhibition of the chlorophyll molecules, brought about by the overreduction of the components of the photosynthetic ETC, was reviewed in Section 2.2.4.4. A second objective motivating this study was to investigate the various routes through which NADPH produced during the light reactions and exported to the cytosol via the various shuttle mechanisms are converted to ATP by the mitochondrial ETC.

For the purposes described above, independent models presented in the previous chapter were integrated using relevant transport reactions. While the first part of this chapter investigates the characteristics of the exchange of ATP and reducing equivalents between chloroplast and cytosol, the latter part examines the effect of mitochondrial metabolism on ATP and redox interactions between chloroplast and cytosol, and the various routes involved in controlling the overreduction of photosynthetic ETC.

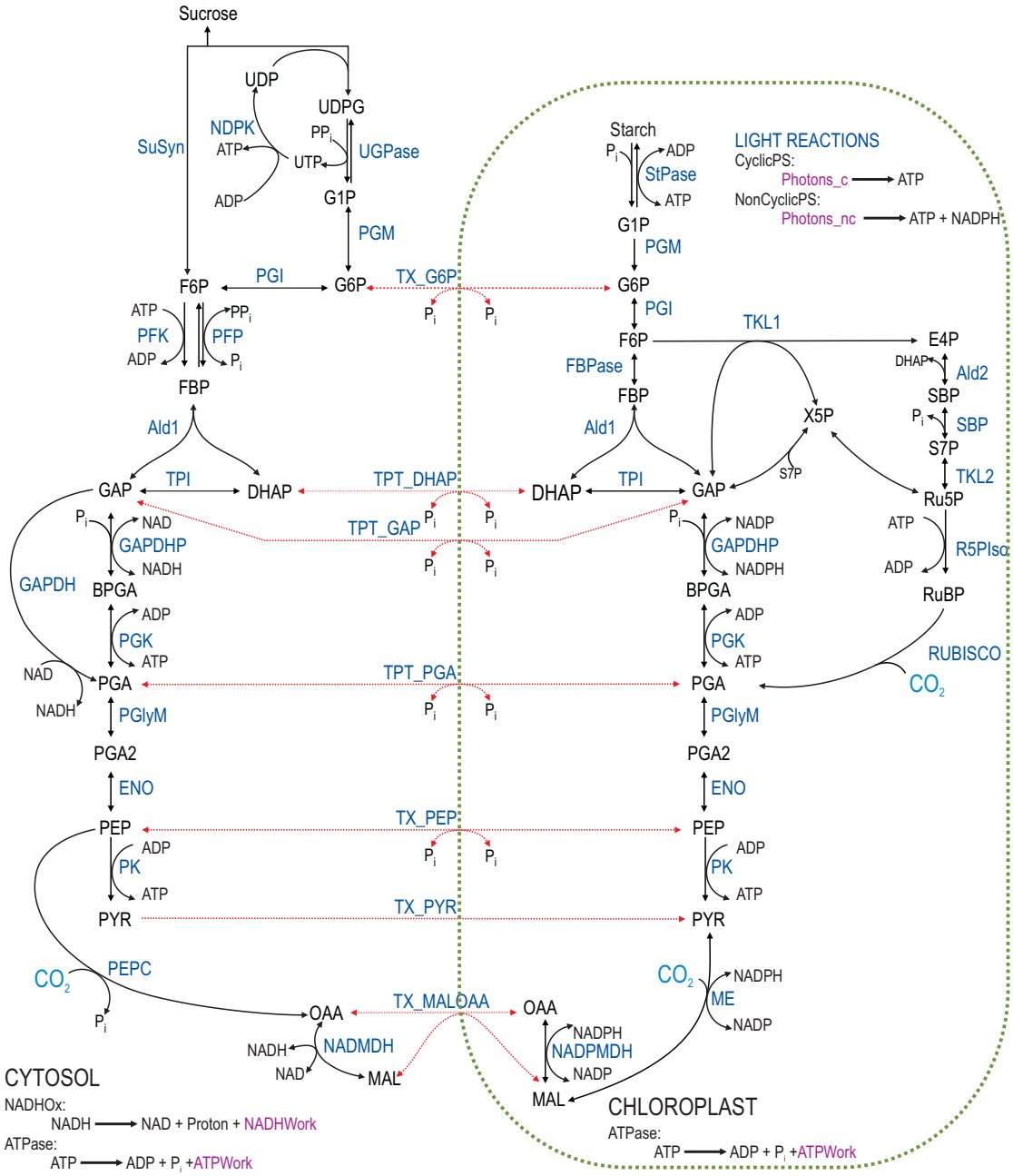


Figure 4.1 – Schematic representation of the integrated model of light reactions, the Calvin cycle and glycolysis representing the energy and redox interactions between chloroplast (green dotted enclosure) and cytosol. Transport reactions between the compartments are shown using red arrows. External metabolites are highlighted in purple. Suffixes '_str' and '_cyt' representing the stromal and cytosolic localisation of the metabolites and reactions are not shown here for clarity, but were included in the model. See text for detailed description of the model definition and List of Abbreviations for metabolite abbreviations.

4.2 Interaction between chloroplast and cytosol

4.2.1 Model Integration

The steady-state stoichiometric models of the Calvin cycle (Section 3.3) and glycolysis (Section 3.4) were integrated with the help of transport reactions that mediate the trans-

Table 4.1 – Enzyme subsets in the combined model of glycolysis and chloroplast metabolism, involving the light reactions and the Calvin cycle in the absence of net CO_2 fixation. See Figure 4.1 for a graphical representation of the reactions involved.

Subset	Reactions	Function
1	NADHOx_cyt NonCyclicPS	Reactions involved in the production and transfer of NADPH
2	PK_str TX_PYR_str PK_cyt	This is an inconsistent ES as PK is irreversible and can be ignored during the analysis of EMs
3	NADMDH_cyt TX_MALOAA_str NADPMDH_str	Reactions of the MAL/OAA shuttle
4	TPI_str TPI_cyt	Conversion of GAP and DHAP
5	PGI_str Ald1_str PFK_cyt TX_G6P_str Ald1_cyt PGI_cyt FBPase_str	Transfer of stromal G6P to cytosol and its conversion to cytosolic GAP and DHAP
6	PGlyM_cyt PGlyM_str Eno_cyt Eno_str	Stromal and cytosolic PGA to PEP
7	PGK_cyt GAPDHP_cyt	Reversible conversion of cytosolic GAP to PGA
8	PGK_str GAPDHP_str	Reversible conversion of stromal GAP to PGA
9	ME_str, Ald2_str Rubisco_str, X5Piso_str SuSyn_cyt, UGPase_cyt PGM_str, NDPK_cyt PGM_cyt, R5Piso_str Ru5Pk_str, TKL1_str PEPC_cyt, PFP_cyt StSynth_str, TKL2_str	Dead reactions Reactions that are not involved in the transfer of ATP and redox equivalents

fer of Calvin cycle intermediates G6P, DHAP, GAP, PGA, PEP, PYR, MAL and OAA from chloroplast to cytosol. ATP and NADPH required for the Calvin cycle to produce these intermediates were supplied to the combined model by incorporating a simplified version of the model of the light reactions described in Section 3.2 (Appendix A). The simplified model contains two independent reactions named 'NonCyclicPS' and 'CyclicPS' representing the overall stoichiometries of the EMs 1 and 2, respectively, of the light reactions shown in Table 3.2. While 'NonCyclicPS' represents the conversion of an external metabolite 'Photon_nc' to stromal ATP and NADPH, 'CyclicPS' represents the formation of stromal ATP from the external metabolite 'Photon_n'. The reason for introducing the simplified model was to simplify the future analysis and interpretation of the combined model. Stoichiometries of the independent models of the light reactions, the Calvin cycle and glycolysis is available in ScrumPy format in

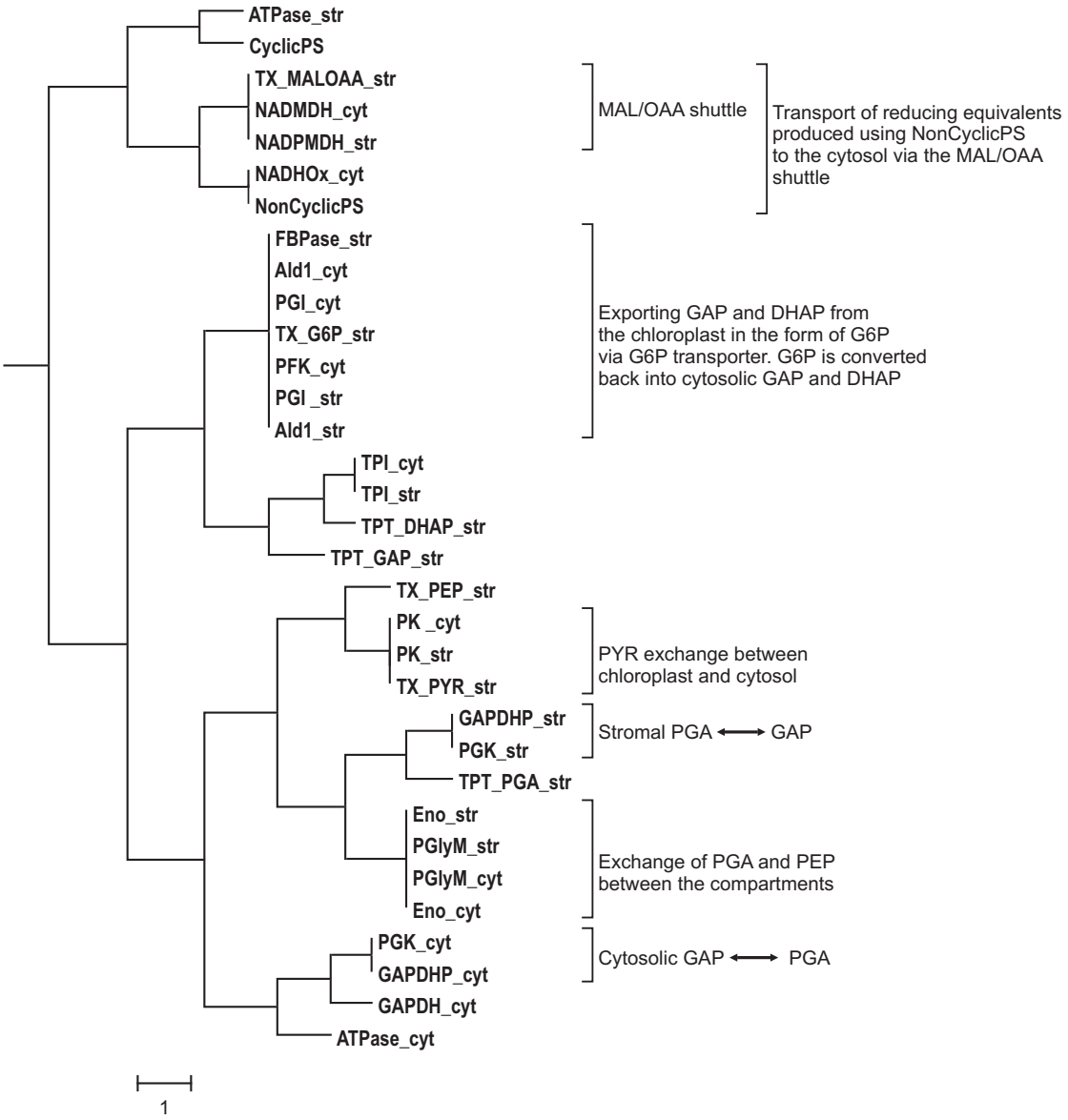


Figure 4.2 – Metabolic tree representing the correlations between fluxes carried by reactions in the combined model of light reactions, the Calvin cycle and glycolysis. All carbon flux in the model was shut off by removing carbon input into the system. See List of Abbreviations for metabolite abbreviations and Appendix A for the stoichiometries of reactions. The scale bar respresents a difference of $\theta^K_{xy} = 1$ rad.

Appendices A, B and C, respectively. An illustration of the combined model is shown in Figure 4.1. The final version of the model contained 49 reactions and 53 metabolites.

Exchange of the intermediates of the Calvin cycle — produced using ATP and NADPH generated during the light reactions — between chloroplast and cytosol forms the key structure of this network. Carbon flux through this model was ascertained using standard stoichiometric model analysis techniques, a dendrogram showing the correlation between fluxes carried by reactions in an extended model (that also contains the TCA cycle reactions) is shown in Figure 4.8. Several modifications were made to this combined model in order to investigate the exchange of ATP and reducing equivalents without any net carbon flux. External metabolites of the independent models

— starch, sucrose and PYR — were now made internal and the dummy metabolites, stromal and cytosolic 'ATPWork' and 'NADWork', were declared external. These dummy metabolites were originally introduced into the model to directly derive net energy and reducing yields from the net stoichiometries of the EMs. Once the carbon flux is shut off and the external metabolites defined, the combined model now represents the exchange of ATP and redox equivalents generated during the light reactions between chloroplast and cytosol.

4.2.2 Model Analysis

ES analysis of the combined model of the light reactions, the Calvin cycle and glycolysis revealed nine subsets composed of more than two reactions and eight involving only a single reaction. A list of reactions constituting the major ESs and their respective functions are shown in Table 4.1. Smaller subsets are not listed here for clarity. The largest subset (ES 9), containing 16 reactions, represented reactions that were not involved in the exchange of ATP and reducing equivalents. These reactions, although integral to the independent models (Sections 3.2, 3.3 and 3.4), do not have any flux associated with them as the net carbon flux in the model was initially shut off. The second largest subset, ES 5, contained seven reactions that were involved in the conversion of stromal DHAP and PGA to G6P, the export of G6P to the cytosol via the G6P transporter (TX_G6P) and its subsequent conversion to cytosolic GAP and PGA at the expense of a molecule of ATP.

A metabolic tree based on reaction correlation coefficients, representing the correlations between fluxes carried by reactions in the model, was constructed from the orthogonal null space of the stoichiometry matrix of the combined model (Figure 4.2). Clusters that correspond to ESs shown in Table 4.1 were identified. The properties of some of these clusters are provided in Figure 4.2.

EM analysis of the model revealed 36 EMs, the overall stoichiometries of which were hierarchically clustered, as shown in the dendrogram in Figure 4.3. Seven distinct clusters were observed in the dendrogram. Clusters 1, 2 and 4 contained EMs that were not involved in the transfer of ATP and/or reducing equivalents across the chloroplast envelope. EMs in Cluster 1 were futile cycles involving DHAP and GAP, and PGA and PEP transporters. While Cluster 2 contained EMs that use ATP and/or NADPH from the stroma to initiate substrate cycles that do not export them to the cytosol, Cluster 4 represented an EM that converted ATP produced from photons into the stroma to stromal ATPWork. These clusters, containing a total of 13 EMs, were removed from the original set of EMs and will not be discussed further. Overall stoichiometries of the remaining 23 EMs were hierarchically clustered and are shown in the dendrogram in Figure 4.4. Clusters 3 and 5 of this dendrogram contained EMs involved in the transfer of ATP

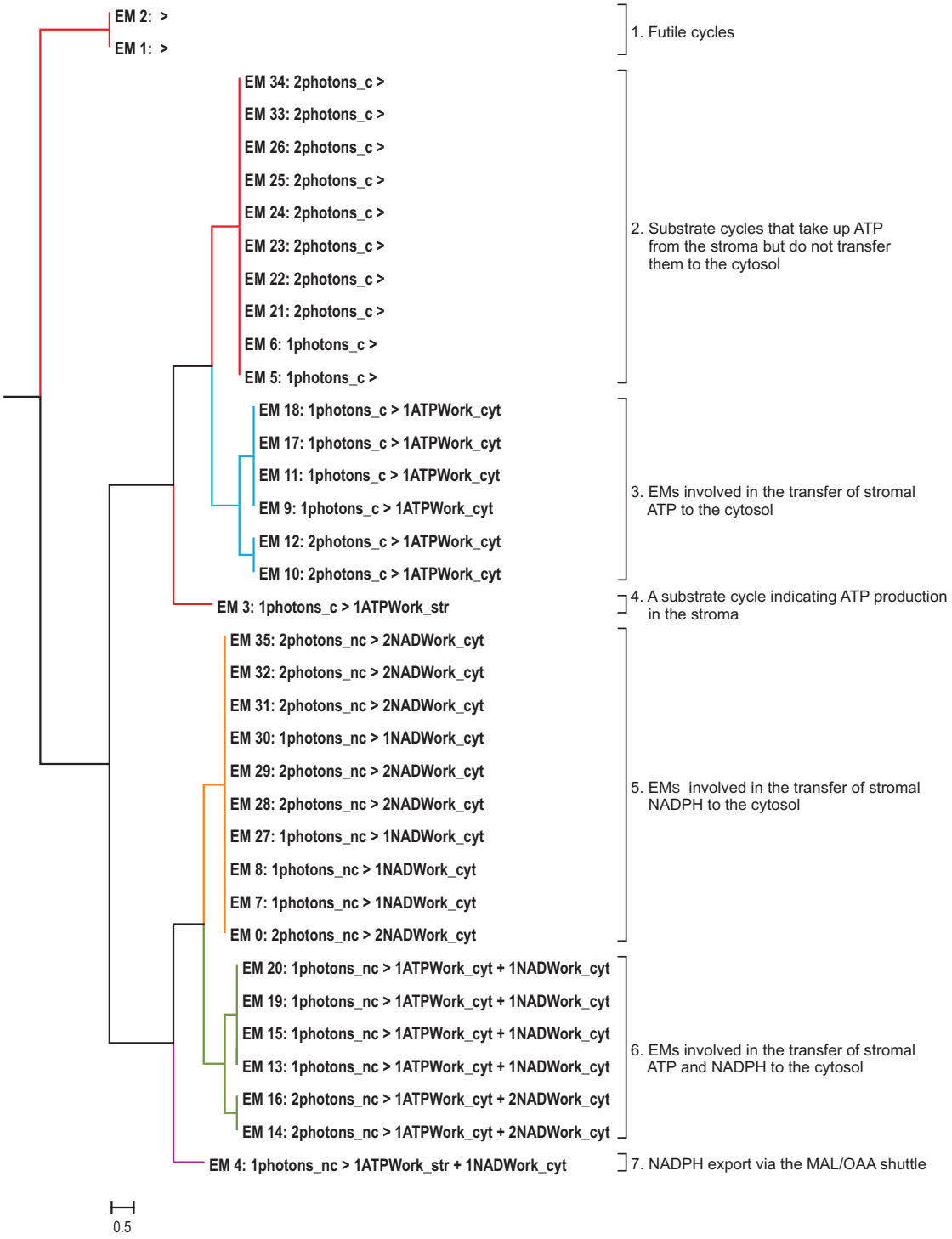


Figure 4.3 – Dendrogram representing the overall stoichiometries of the entire set of EMs generated from the combined model of the light reactions, the Calvin cycle and glycolysis, clustered by angle based on their net external metabolite usage. Blue, orange, green and purple clusters contain EMs that mediate the transfer of ATP and/or reducing equivalents between chloroplast and cytosol. EMs in red clusters are not involved in either energy or redox exchange. '_str' and '_cyt' indicate the localisation of the external metabolites in stroma and cytosol, respectively. See List of Abbreviations for metabolite abbreviations. The scale bar respresents a difference of $\theta^{K}_{xy} = 0.5$ rad.

and NADPH from the chloroplast to the cytosol, respectively. Cluster 7 represented an EM that uses the MAL/OAA shuttle to export NADPH to the cytosol. EMs capable of

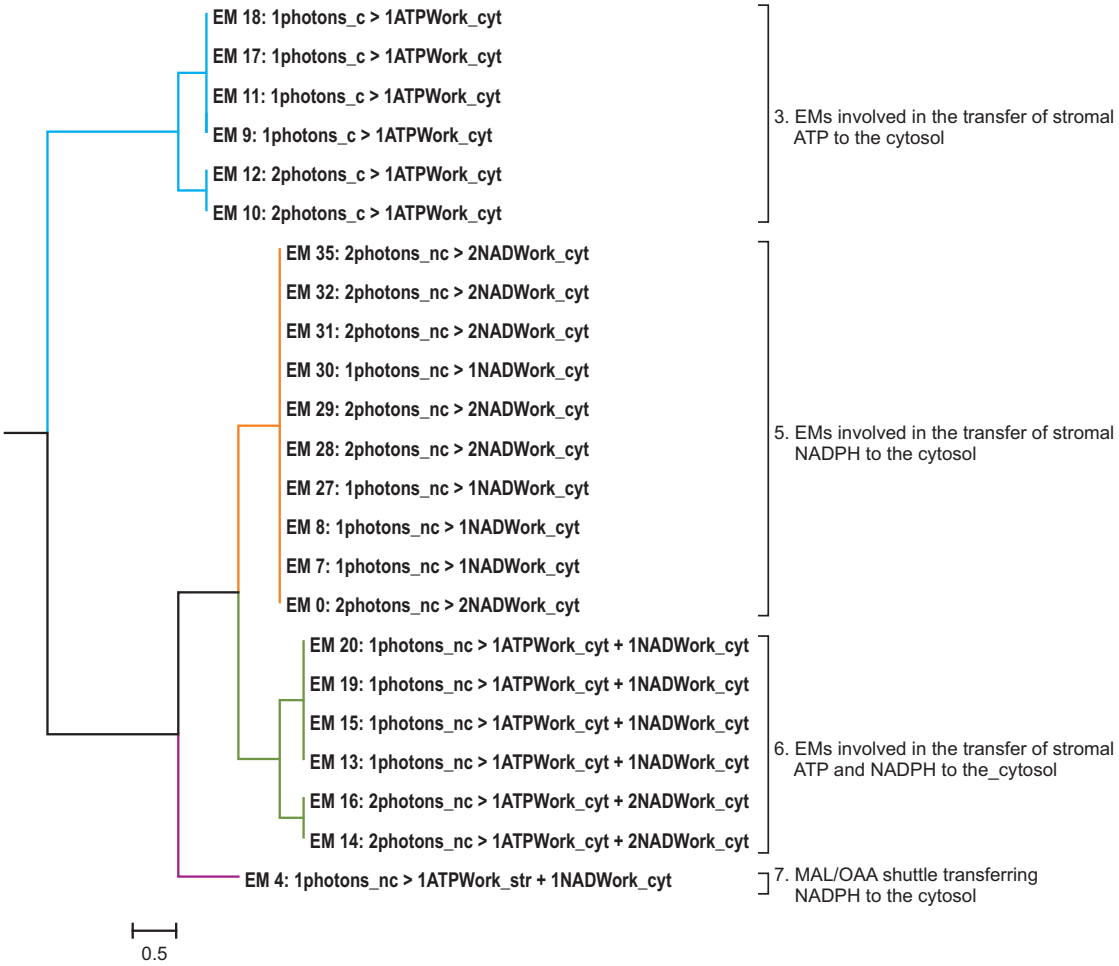


Figure 4.4 – Dendrogram representing the overall stoichiometries of the 23 EMs obtained after removing the EMs that are not involved in the transfer of ATP and/or reducing equivalents. Blue, orange, green and purple clusters contain EMs that mediate the transfer of ATP and/or reducing equivalents between chloroplast and cytosol. '_str' and '_cyt' indicate the localisation of the external metabolites in stroma and cytosol, respectively. See List of Abbreviations for metabolite abbreviations. The scale bar respresents a difference of $\theta_{xy}^{K} = 0.5$ rad.

exporting both ATP and NADPH from the stroma were grouped in Cluster 6. A heatmap representing reactions participating in the final set of EMs was constructed from the EMs reaction matrix (Figure 4.5). Rows of this matrix were hierarchically clustered based on the overall stoichiometries of the EMs and columns were sorted based on the order of the leaves on the reaction correlation tree shown in Figure 4.2.

4.2.3 Discussion

The major objective of performing ES analysis of the model was to identify reactions that operate in fixed flux proportions. Results from the ESs analysis, shown in Table 4.1, revealed some very important properties of the integrated model. Reactions in ES 9, identified as dead reactions, are obviously not involved in the exchange of either ATP or reducing equivalents between the compartments. Note that the objective of performing

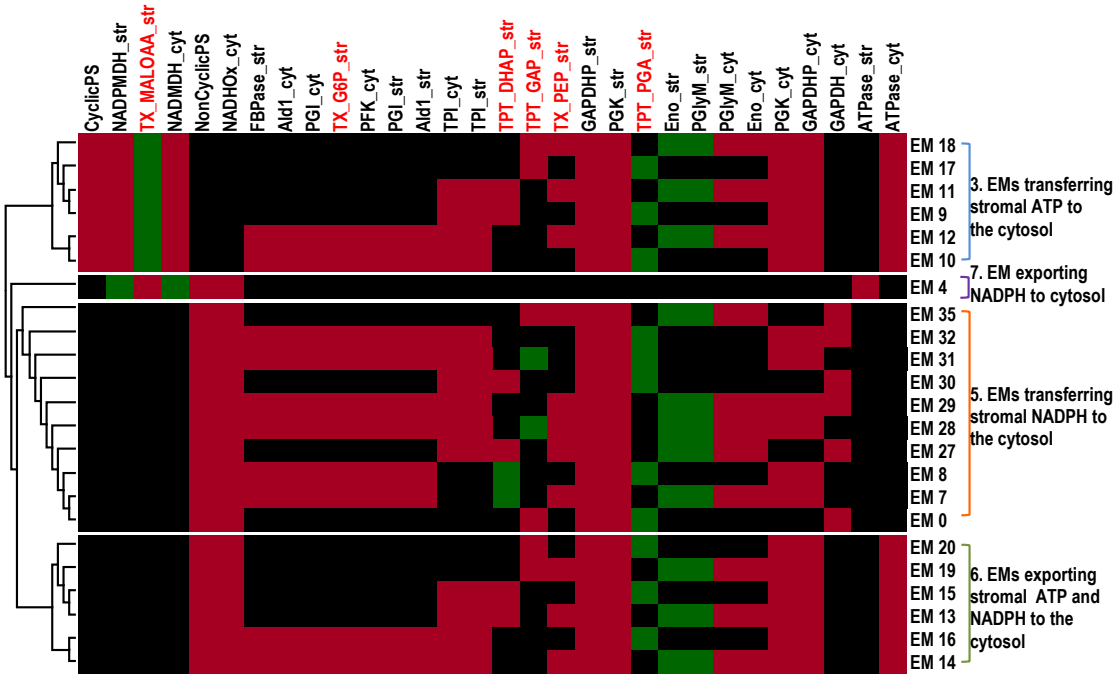


Figure 4.5 – Heatmap representing reactions participating in the elementary modes of the combined model of glycolysis and chloroplast metabolism. Rows were clustered based on the overall stoichiometries of EMs and the columns were clustered based on reaction correlation coefficients that represent the correlations between fluxes carried by reactions in the model. Each EM represents a potential route involved in the transfer of ATP and/or reducing equivalents from chloroplast to cytosol. Blue, orange, green and purple clusters relate to those in the dendrogram in Figure 4.4. Transport reactions involved in these modes are highlighted in red text. '_str' and '_cyt' indicates the localisation of the external metabolites in the stroma and cytosol, respectively. The heatmap has been coloured according to the stoichiometric coefficients of reactions in an EM i.e. $= 0$ (black), < 0 (green) and > 0 (red) (i.e.. red and green indicate forward and backward reactions, respectively. Black represents non-participation of a reaction in an EM).

the structural analysis of the integrated model was to investigate the major routes through which ATP and reducing equivalents may be exported from chloroplast to cytosol without any net carbon flux. Reactions in ES 9 reiterates the absence of any carbon flux in the model. ES 5, containing reactions involved in the conversion of stromal GAP and DHAP to cytosolic GAP and DHAP via the G6P transporter, represents the only means by which the G6P transporter can exchange ATP and redox equivalents across the chloroplast envelope. Net stoichiometry of this ES involves utilisation of cytosolic ATP, so when it occurs as part of EMs involved in chloroplast-cytosol transfer, it can only achieve net transfer of NADPH as ATP from the chloroplast is consumed.

The reasons for performing EM analysis on the integrated model of the light reactions, the Calvin cycle and glycolysis were twofold: to identify potential routes through which ATP and reducing equivalents are transferred across the chloroplast envelope without any net carbon flux and to interrogate the properties and behaviours of the major transporters involved in this transfer. From the 36 possible EMs obtained from the model, 23 distinct routes responsible for ATP and redox transfer

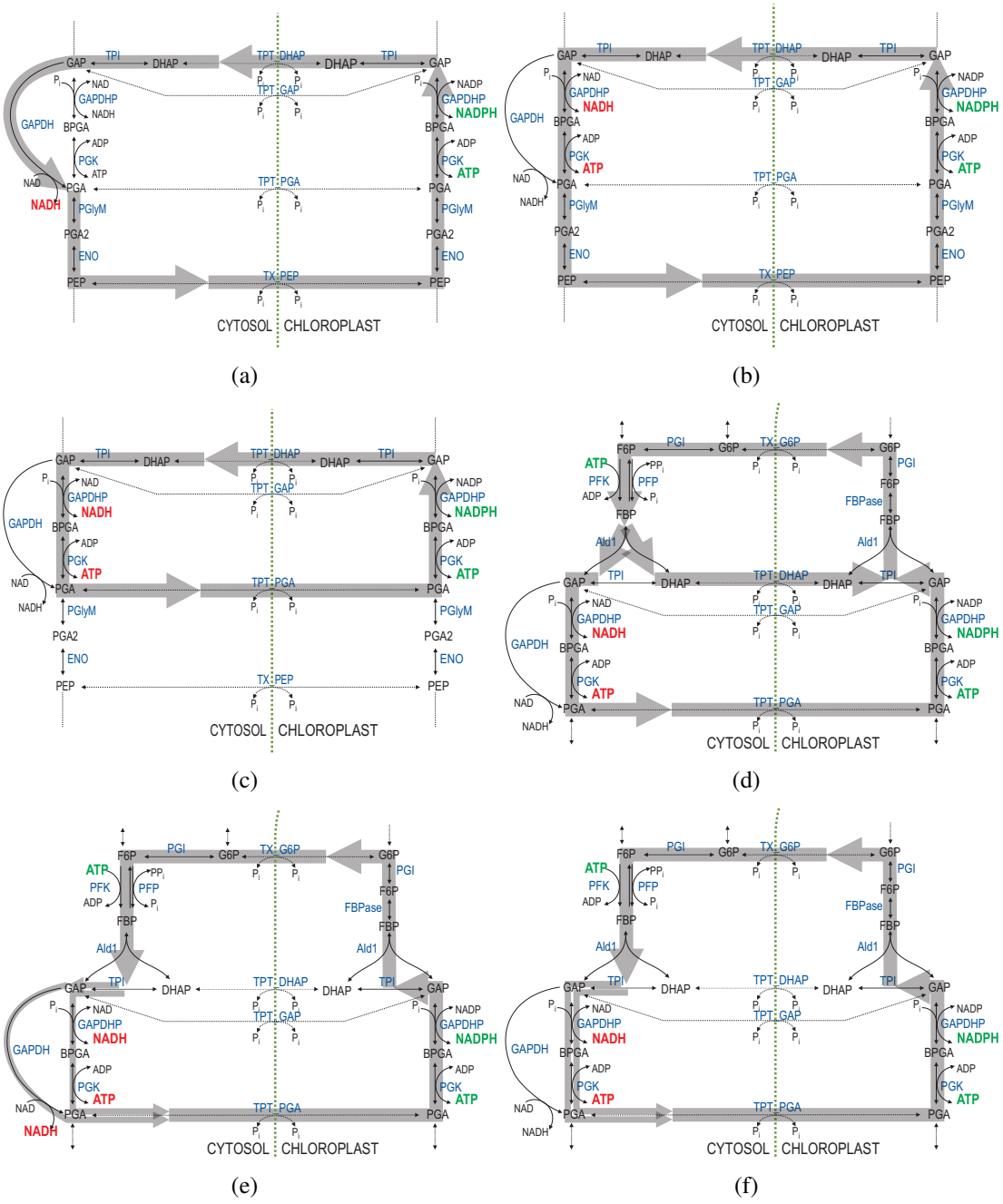


Figure 4.6 – A representative set of EMs of the integrated model of light reactions, the Calvin cycle and glycolysis involved in the transfer of ATP and/or reducing equivalents from chloroplast to cytosol. (a), (b), (c), (d), (e) and (f) represent EMs 35, 20, 15, 8, 32 and 16, respectively. Overall stoichiometries of these EMs are shown Figure 4.4. EMs and their flux are indicated by the grey overlay. Energy and reducing equivalents consumed and produced in each of these EMs are coloured green and red, respectively. While (a) and (b) indicate the role of PEP transporter in the transfer of ATP and reducing equivalents, (c), (d), (e) and (f) illustrate the role of G6P transporter. See text for more details on these EMs.

were identified. When the overall stoichiometries of these EMs were hierarchically clustered, four distinct groups of EMs were identified (Figure 4.4). The first cluster, Cluster 3, containing six EMs was found to be involved in the transfer of ATP. Overall stoichiometries (Figure 4.4) and participating reactions (Figure 4.5) indicate that the

ATP transferred by these EMs came from the cyclic branch of the light reactions. However, the presence of ESs 7 and 8 containing cytosolic and stromal PGK and GAPDHP suggests that the ATP transfer through these modes must also involve coupled NADPH export from chloroplast. This explains the reason for having the MAL/OAA shuttle (ES 3) associated with these EMs. Reducing equivalents produced in the cytosol by GAPDHP are transferred into the stroma via the MAL/OAA shuttle, to provide the EMs with the NADPH required to transfer ATP into the cytosol. For these reasons, as these EMs are potential routes for ATP transfer into the cytosol, it can be concluded that cyclic photophosphorylation cannot mediate ATP transfer into the cytosol on its own.

Clusters 5 and 7 contain EMs involved in the transfer of reducing equivalents produced during non-cyclic photophosphorylation across the chloroplast envelope. While Cluster 7 contains only one EM that uses the MAL/OAA shuttle to do this, Cluster 5 contains ten EMs that use a combination of different transporters other than the MAL/OAA shuttle. Two main groups of EMs can be observed in Cluster 5. The first group of EMs employ stromal PGK and GAPDHP, and cytosolic GAPDH to transfer reducing equivalents into the cytosol. A representative EM of this cluster, EM 35, is illustrated in Figure 4.6(a). The second group imports stromal G6P produced from PGA into the cytosol via the G6P transporter, and converts it into cytosolic FBP with the help of PGI and PFK reactions. FBP is then broken down into two molecules of GAP by the cytosolic Ald1 and TPI reactions. EMs in this cluster show that both these molecules can follow independent routes from this point. In the case of the EMs 8, 7, 28 and 31, while one molecule of GAP moves back into the chloroplast via the TPT_GAP or TPT_DHAP transporter the other gets converted to PGA either via the PGK and GAPDHP reactions or the GAPDH reaction. A representative EM, EM 8, is illustrated in Figure 4.6(d). In EMs 29 and 32, however, both molecules of GAP are converted to PGA - one using GAPDHP and PGK reactions and the other using the GAPDH reaction. See Figure 4.6(e) for a representative example. Consumption and production of a molecule of ATP by cytosolic PFK and PGK, respectively, in these EMs means that there is only a net import of reducing equivalents into the cytosol. EMs in this group exemplify the role of the G6P transporter in transferring reducing equivalents across the chloroplast envelope. However, EMs 16 and 14 of Cluster 6 are capable of a net export two molecules of NADPH and a molecule of ATP into the cytosol. In the case of these EMs both molecules of GAP generated from FBP are converted to PGA by the cytosolic reactions GAPDHP and PGK, forming two molecules of NADPH and ATP. One molecule of this ATP is used up by the PFK reaction while converting F6P to FBP. See Figure 4.6 (f) for an illustration of EM 16. The other three EMs in Cluster 6 use stromal and cytosolic PGK and GAPDHP to export both ATP and NADPH from the chloroplast (Figures 4.6(c) and (b)).

From the heatmap in Figure 4.5 it is evident that all EMs except the one in Cluster 3 use stromal PGK and GAPDHP to initiate the export of ATP and redox equivalents.

Along with these two stromal reactions the cytosolic PGK and GAPDHP form the most important enzymes required to facilitate the exchange of ATP and NADPH between the two compartments. The heatmap also shows that the flux through the PGA transporter is directed into the chloroplast in all EMs that use it. Similarly, the direction of flux through stromal reactions PGlyM and Eno means that the flux through PEP transporter is directed into the chloroplast. These two transport reactions replenish the stroma with the metabolites PGA and PEP required to initiate the transfer of ATP and reducing equivalents. DHAP and GAP transporters, on the other hand, are involved in the exchange of triose-phosphates across the chloroplast envelope. The resulting pathway is mediated by coupled reactions GAPDHP, PGK and TPI running anti-parallel in the stroma and cytosol. These transporters are involved in 15 EMs. In four EMs involving the G6P transporter, the GAP and DHAP transporters mediate the influx of triose-phosphates into the stroma. The role of the G6P transporter in the model was described earlier. The above observations suggest that apart from the triose-phosphate, PGA and MAL/OAA transporters described earlier by Heineke *et al.* [148], two additional transporters — G6P and PEP transporters — are potentially involved in the exchange of ATP and reducing equivalents across the chloroplast envelope. However, more sophisticated techniques such as FBA and kinetic modelling must be employed to ascertain the exact nature of their role. Another important outcome of this analysis is that the observations highlight the role of the chloroplast as the source and target of redox regulations in the plant cell. Integrating mitochondrial metabolism to the existing model might enable us to investigate this further.

4.3 Energy and redox interactions between chloroplast, cytosol and mitochondria

4.3.1 Model extension

An integrated steady state structural model containing light reactions, the Calvin cycle and glycolysis representing the exchange of ATP and reducing equivalents between chloroplast and cytosol was described in the previous section. This model was extended to include mitochondrial metabolism by integrating the stoichiometric model of the TCA cycle described in Section 3.5. The models were combined with the help of MAL/OAA and PYR transporters that mediate the exchange of metabolites between mitochondria and cytosol. Stoichiometries of the independent models of light reactions, the Calvin cycle, glycolysis and TCA cycle in ScrumPy format is available in Appendices A, B, C and D, respectively. An illustration of the integrated model is shown in Figure 4.7. The final version of the model contains 63 reactions and 72 metabolites.

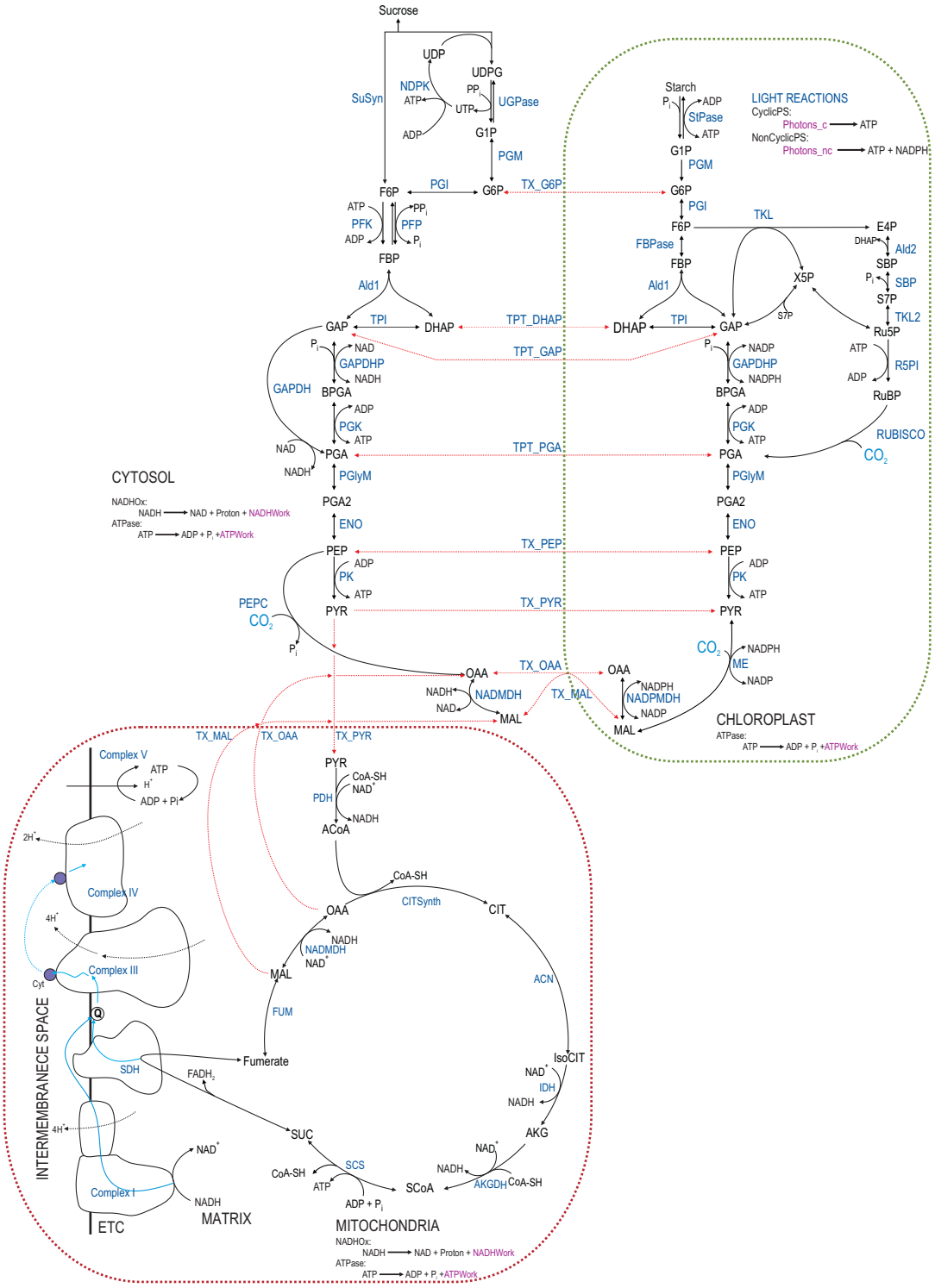


Figure 4.7 – Schematic representation of the integrated model of light reactions, the Calvin cycle, glycolysis and the TCA cycle representing the energy and redox interactions between chloroplast (green dotted enclosure), cytosol and mitochondria (red dotted enclosure). Reaction of the oxidative phosphorylation were included in the model definition to simplify future analysis and interpretation of the model. Transport reactions between the compartments are highlighted in red arrows. External metabolites are highlighted in purple. Suffixes '_str', '_cyt' and '_mit' representing the stromal, cytosolic and mitochondrial localisation of the metabolites and reactions are not shown here for clarity, but were included in the model. See text for detailed description of the model definition and List of Abbreviations for metabolite abbreviations.

One objective of constructing this model was to investigate the major routes through which ATP and reducing equivalents are exchanged between chloroplast, cytosol and mitochondria without any net carbon flux. However, the extended model was tested using standard stoichiometric model analysis techniques to ascertain the existence of carbon flux. A dendrogram representing the correlation between fluxes carried by reactions in the model with carbon flux is shown in Figure 4.8. In order to restrict carbon flux in the extended model, PYR, MAL and OAA, which had been considered external in the independent model of TCA cycle and oxidative phosphorylation, were now made internal. Note that the carbon flux through the initial model of light reactions, the Calvin cycle and glycolysis was already shut off.

The reactions involved in oxidative phosphorylation were not included in the definition of the extended model for two reasons. Firstly, it simplified the model definition and future analysis and interpretation of the model. Secondly, it was shown during the analysis of the reactions participating in EM 4 in Figure 3.13 that the reactions of the oxidative phosphorylation can mediate direct oxidation of mitochondrial NADH to yield ATP without the involvement of any carbon flux. If follows from this that any NADH formed in the mitochondrial matrix is ultimately converted to mitochondrial ATP. For this reason, a new dummy metabolite, 'NADHWork', was introduced into the mitochondrial model to derive net reducing yields from the net stoichiometries of the EMs. Compartmentation of metabolites and reactions were maintained by using the suffixes '_str', '_cyt' and '_mit' representing stroma, cytosol and mitochondria, respectively. The extended model now represented the exchange of ATP and reducing equivalents produced during light reactions between chloroplast, cytosol and mitochondria without any net carbon flux.

4.3.2 Model Analysis

Reactions involved in the major ESs of the initial model, constituting light reactions, the Calvin cycle and glycolysis, were shown in Table 4.1 and the characteristics of these ESs were described in Section 4.2.2. ESs analysis of the extended model showed that all subsets in the initial model, both major and minor, were conserved. In addition to this, two novel characteristics were observed. Firstly, the total number of reactions in the largest ES of the initial model, ES 9, rose from 16 to 27. This subset, already containing reactions of the stromal and cytosolic metabolism that were considered dead, now included reactions of the TCA cycle that were not involved in the transfer of ATP or reducing equivalents. The amended list of reactions in ES 9 is shown in Table 4.2. Note that all reactions in the TCA cycle are dead in the absence of carbon flux in the model. Secondly, a new ES (ES 10 in Table 4.2) containing reactions mediating the mitochondrial MAL/OAA shuttle was identified. This ES represents the only route through which reducing equivalents can enter or leave mitochondrial matrix.

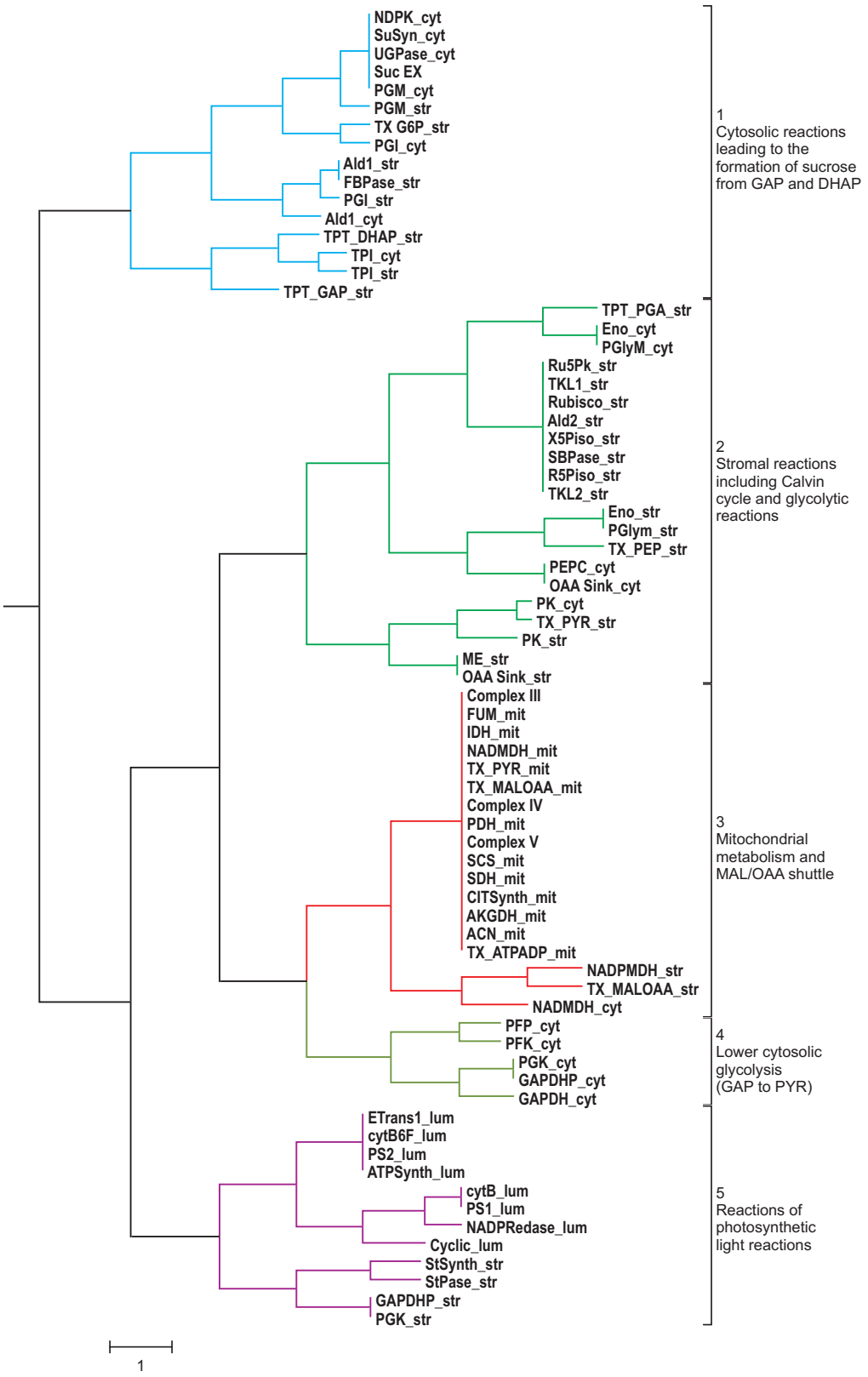


Figure 4.8 – Dendrogram representing the correlation between fluxes carried by reactions in the combined model of light reactions, the Calvin cycle, glycolysis and the TCA cycle in the presence of carbon flux. Five distinct clusters were observed, the general characteristics of each one of them is presented. '_lum', '_str', '_cyt' and '_mit' indicates localisation of metabolites in the lumen, stroma, cytosol and mitochondria, respectively. See the List of Abbreviations for reaction name abbreviations. Scale bar respresents a difference of θ^{K}_{xy} = 1rad.

Table 4.2 – ESs obtained from the extended model. ESs of the initial model comprising light reactions, the Calvin cycle and glycolysis were shown in Table 4.1. See Figure 4.7 for a graphical representation of the reactions involved.

Subset	Reactions	Function
9	ME_str, Ald2_str Rubisco_str, X5Piso_str SuSyn_cyt, UGPase_cyt PGM_str, NDPK_cyt PGM_cyt, R5Piso_str Ru5Pk_str, TKL1_str PEPC_cyt, PFP_cyt StSynth_str, TKL2_str AKGDH_mit, TX_PYR_mit Fumarase_mit, SCS_mit ACOase_mit, CITSynth_mit IDH_mit, PDH_mit, SDH_mit PEPC_cyt, NDPK_cyt	Dead reactions Reactions that are not involved in the transfer of ATP and redox equivalents
10	NADMDH_mit TX_MALOAA_mit NADMDH_cyt	Reactions of the mitochondrial MAL/OAA shuttle

A dendrogram representing the correlation between fluxes carried by reactions in the extended model was constructed after removing those reactions that are not involved in the exchange of ATP and reducing equivalents (Figure 4.9). Clusters on this tree corresponded to those on the tree representing the initial model shown in Figure 4.2. However, a new cluster representing the reactions mediating mitochondrial MAL/OAA shuttle (ES 10) was identified.

EM analysis of the extended model revealed 53 EMs, the overall stoichiometries of which were hierarchically clustered as shown in the dendrogram in Figure 4.10. It was found that all EMs of the initial model of light reactions, the Calvin cycle and glycolysis were conserved in the new set of EMs, i.e. clusters 1-7 containing EMs of the initial model (Figure 4.3) were also present in the dendrogram representing EMs of the extended model (Figure 4.10). Furthermore, clusters 1, 2 and 4 in Figure 4.3, containing 13 EMs that were not involved in ATP and redox exchange, were also present in the new dendrogram. EMs in these clusters were removed from the original set of EMs of the extended model and will not be discussed further. Properties of the remaining 23 EMs in the conserved clusters 3, 5, 6 and 7 of the initial model were described in Section 4.2.2.

In addition to the conserved clusters described above, 3 new clusters — clusters 8, 9 and 10 — containing a total of 17 EMs were found in the dendrogram representing the overall stoichiometries of the EMs of the extended model. Clusters 8 and 9 contained EMs capable of transferring reducing equivalents generated in the stroma into the mitochondrial matrix. EMs in cluster 9 were also capable of transferring stromal ATP into the cytosol. The EM in Cluster 10, however, was involved in transferring stromal reducing equivalents into the mitochondrial matrix with the help of the stromal and mitochondrial MAL/OAA shuttle. A dendrogram containing the 40 EMs that were involved in ATP and redox exchange was constructed and is shown in Figure 4.11.

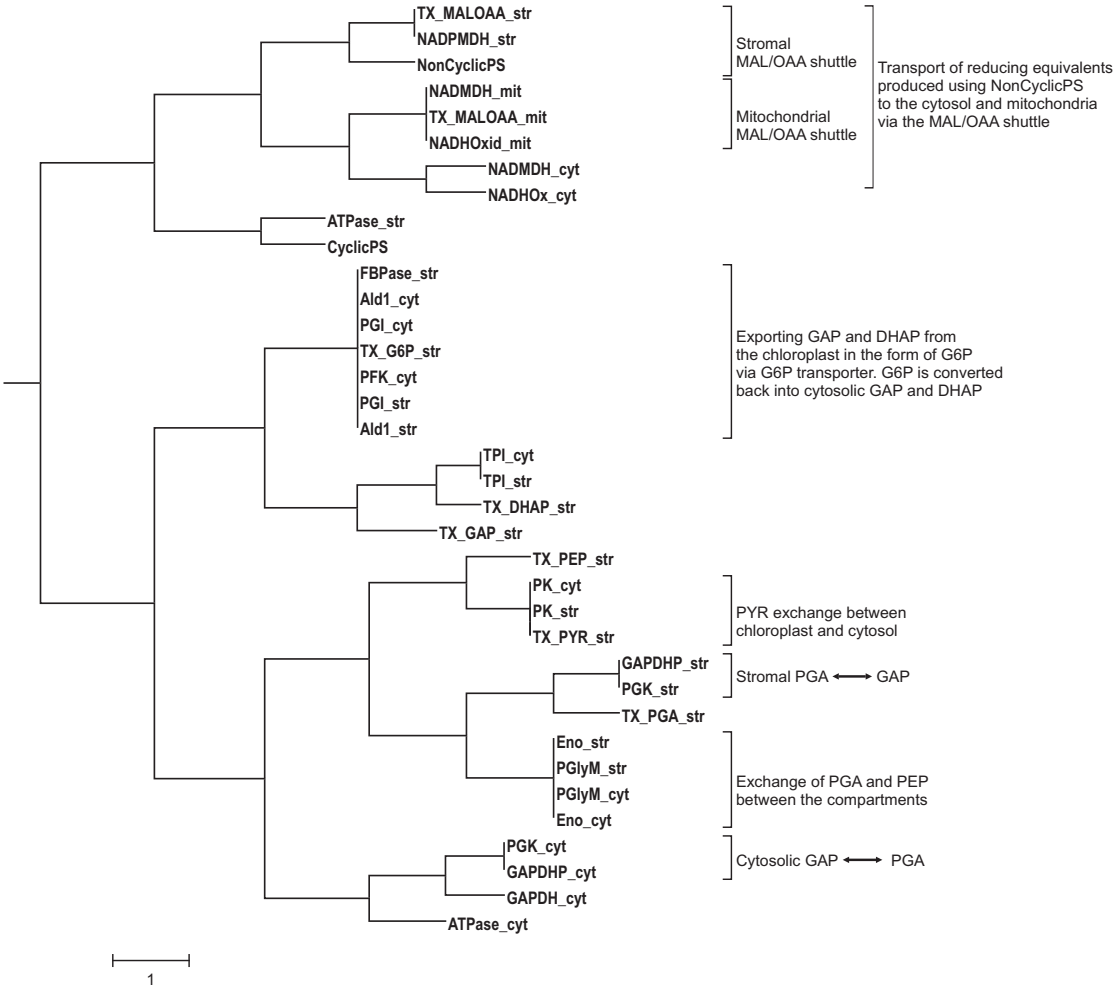


Figure 4.9 – Metabolic tree representing the correlation between fluxes carried by reactions in the extended model composed of light reactions, the Calvin cycle, glycolysis and reactions of the TCA cycle. All carbon flux in the model was shut off by removing carbon input into the system. See List of Abbreviations for metabolite abbreviations and Appendix A for the stoichiometries of reactions. The scale bar respresents a difference of $\theta^K_{xy} = 1$ rad.

A heatmap representing the reactions participating in these EMs was constructed from the EMs reaction matrix (Figure 4.12). Rows of this matrix were clustered based on the overall stoichiometries of the EMs and columns were sorted based on the order of leaves on the reaction correlation tree shown in Figure 4.9. Properties of the reactions participating in the EMs of the initial model shown in this dendrogram were described in Section 4.2 and will not be discussed further.

4.3.3 Discussion

The major objective of performing structural analysis of the model was to investigate the major routes through which ATP and reducing equivalents can be exchanged between chloroplast, cytosol and mitochondria without any net carbon flux. The absence of carbon flux in the model was confirmed by the presence of several key reactions in ES 9 that were identified as dead reactions (Table 4.2). Of these, a reaction of particular

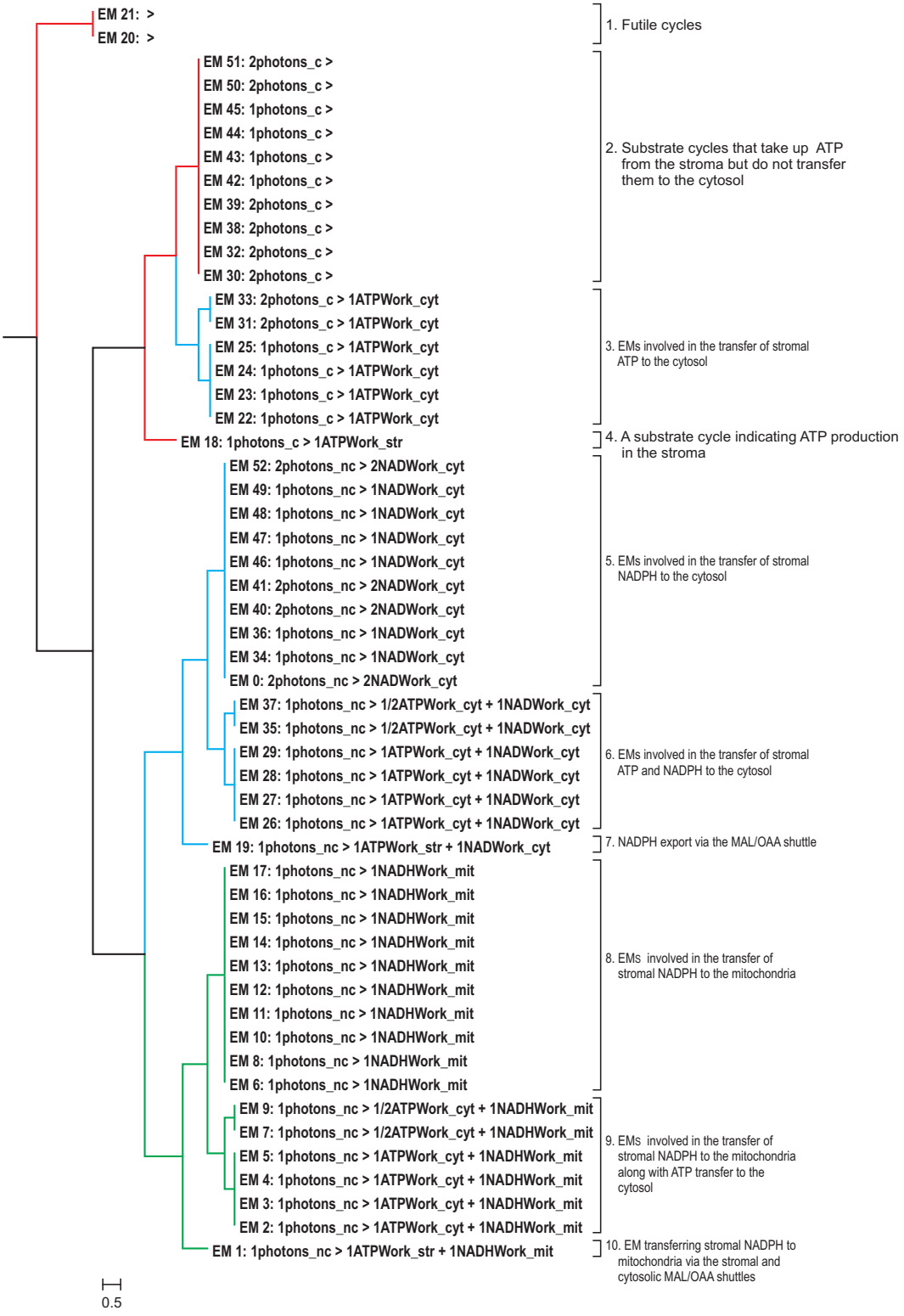


Figure 4.10 – Dendrogram representing the overall stoichiometries of the 53 EMs mediating the exchange of ATP and reducing equivalents in the extended model. Red and blue clusters indicate EMs mediating interaction between chloroplast and cytosol that are conserved from the initial model. Green clusters indicate EMs mediating interaction between chloroplast, cytosol and mitochondria. '_str', '_cyt' and '_mit' indicate the localisation of the external metabolites in stroma, cytosol and mitochondria, respectively. See List of Abbreviations for metabolite abbreviations. The scale bar respresents a difference of $\theta^K_{xy} = 0.5$ rad.

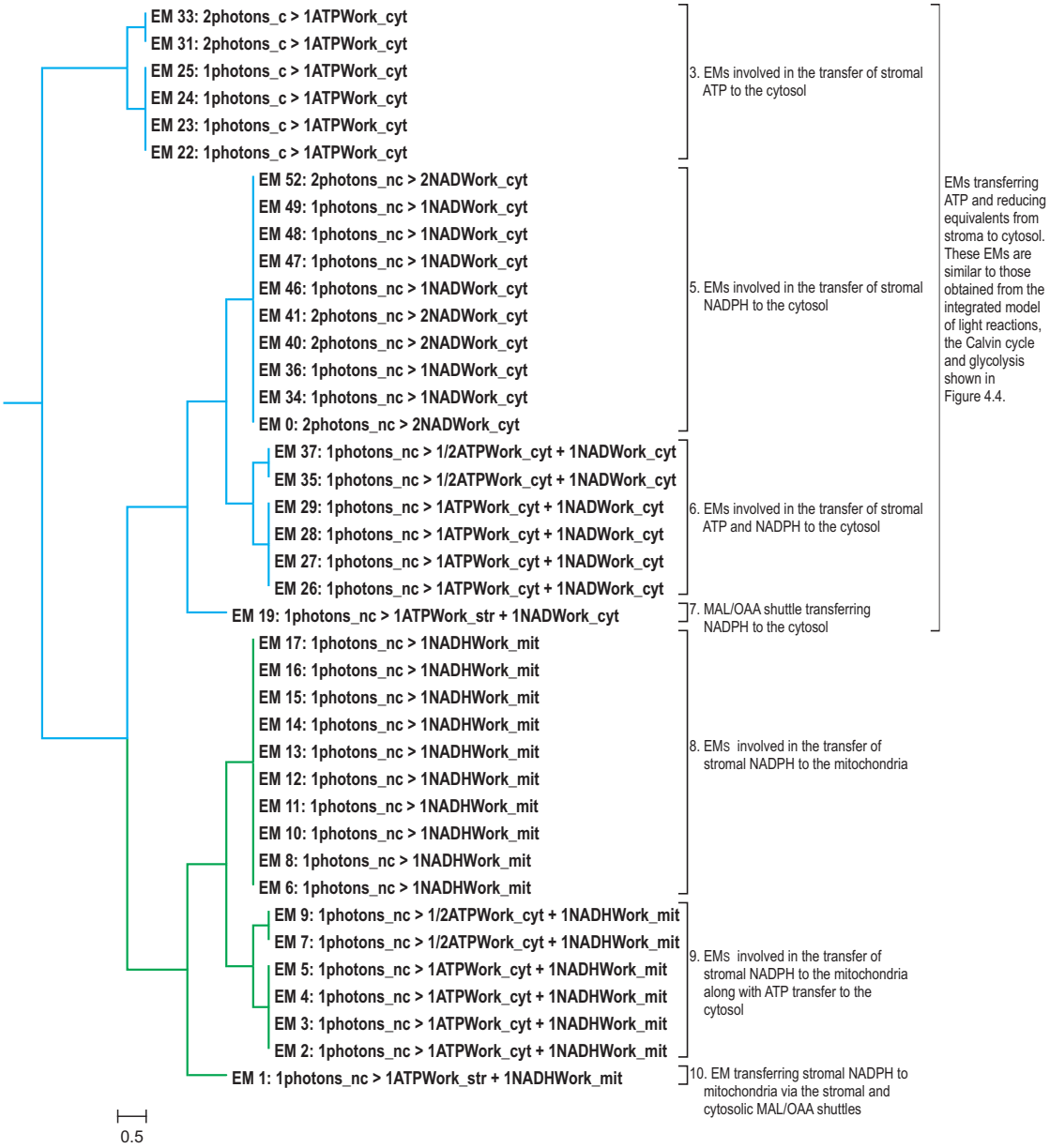


Figure 4.11 – Dendrogram representing the overall stoichiometries of the 40 EMs obtained from the extended model after removing the EMs that were not involved in the transfer of ATP and/or reducing equivalents. Blue clusters contain EMs of the initial model (Figure 4.4) that are conserved in the extended model. Green clusters represent the new set of EMs that are involved in the exchange of ATP and reducing equivalents between chloroplast, cytosol and mitochondria. '_str', '_cyt' and '_mit' indicate the localisation of the external metabolites in stroma, cytosol and mitochondria, respectively. See List of Abbreviations for metabolite abbreviations. The scale bar respresents a difference of $\theta^{K}_{xy} = 0.5$ rad.

interest is the mitochondrial SDH. Having no flux through this reaction means that the model can no longer reduce mitochondrial quinones that can mediate electron transport leading to the production of ATP. Consequently, the only route through which ATP can be generated in the mitochondrial matrix in the absence of carbon flux is via the oxidation of NADH by Complex I. Protons are pumped into the IMS during this process and the resulting proton gradient lead to ATP synthesis. It follows from this

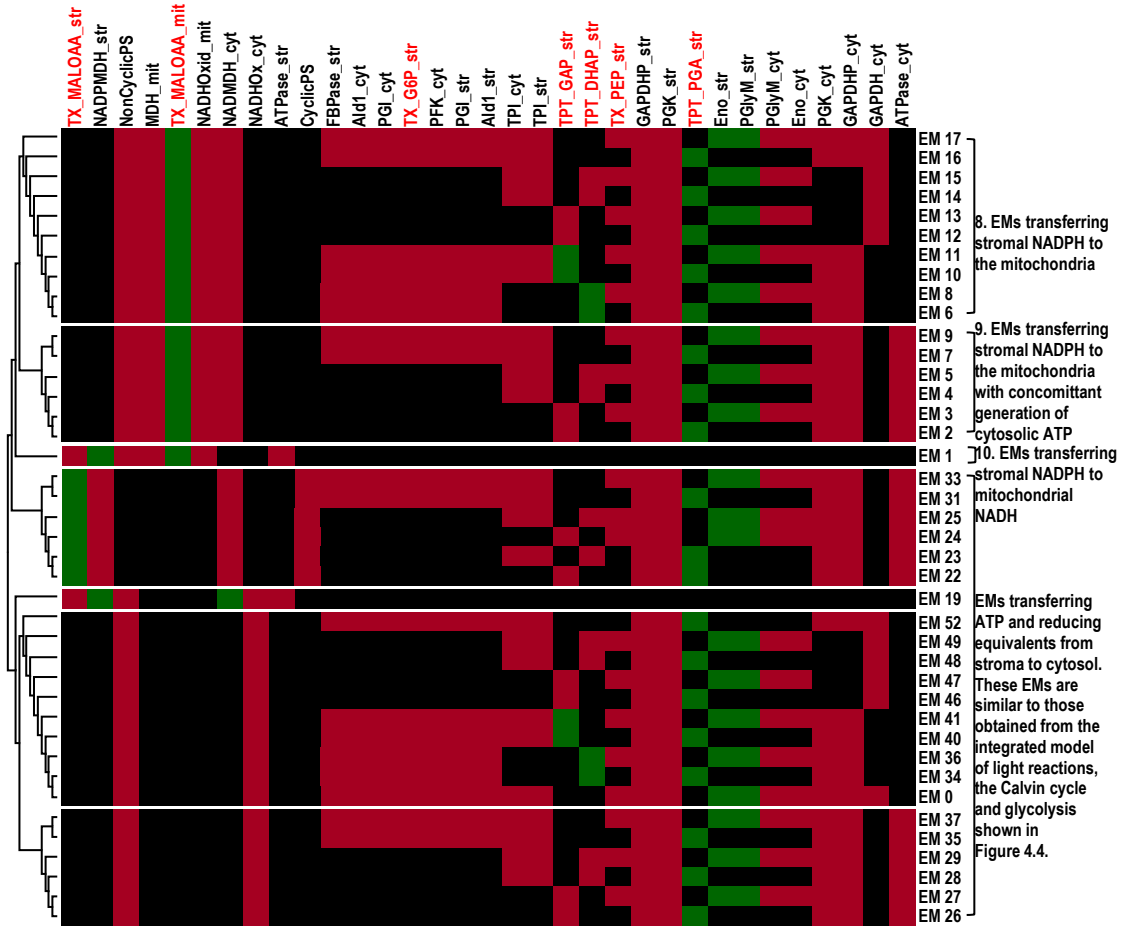


Figure 4.12 – Heatmap representing reactions participating in the elementary modes of the extended model containing light reactions, the Calvin cycle, glycolysis and TCA cycle. Each EM represent a potential route involved in the exchange of ATP and/or reducing equivalents between chloroplast, cytosol and mitochondria. Transport reactions involved in these modes are highlighted in red text. '_str' and '_cyt' indicates the localisation of the external metabolites in the stroma and cytosol, respectively. Heatmap has been coloured according to the stoichiometric coefficients of reactions in an EM i.e. $= 0$ (black), < 0 (green) and > 0 (red) (i.e.. red and green indicate forward and backward reactions, respectively. Black represent non-participation of a reaction in an EM).

that all EMs of the model that can import NADH into the mitochondrial matrix can only use Complex I to produce ATP. ES 10, containing the reactions of the mitochondrial MAL/OAA shuttle, contains the reactions through which reducing equivalents can enter or leave the mitochondrial matrix. The role of this shuttle mechanism in transferring reducing equivalents across the mitochondrial membrane is evident in the heatmap shown in Figure 4.12. Other shuttle mechanisms, such as the MAL/aspartate shuttle, were not considered in the current study as including them would take the model outside the scope of this thesis.

Clusters in the dendrogram representing correlation between fluxes carried by reactions in the initial model were conserved in the dendrogram representing the extended model (Figure 4.9). From the metabolic tree it is evident that the flux through stromal and mitochondrial MAL/OAA shuttles strongly correlate with the flux through

NonCyclicPS. Along with the EMs obtained from the model, it indicates that most reducing equivalents imported into the cytosol from the stroma are taken up by the mitochondria to produce ATP. However, in a live plant cell, cytosolic NADH is used by other metabolic pathways such as nitrate assimilation and some is exported into the peroxisomes. Nevertheless, it can be concluded that excess NADH in the cytosol is exported into the mitochondria along the routes identified by the EMs.

From the dendrograms representing the overall stoichiometries of the EMs of the initial model (Figure 4.3) and the extended model (Figure 4.10) it can be seen that EMs of the former are conserved in the latter. This agrees with the recent observation that modes of a subsystem exist as modes of the enlarged system [183]. Here, reactions in the subsystem containing light reactions, the Calvin cycle and glycolysis were extended by including reactions of the TCA cycle to form the enlarged system. Properties of the EMs of the initial model were discussed in Section 4.2 and were found to be consistent with that of the conserved set of EMs obtained from the extended model. Furthermore, based on the participation of reactions in the EMs of the initial and extended model it can be assumed that the EMs of the subsystem and the enlarged system share some common traits, such as the involvement of stromal GAPDHP and PGK in all EMs and the direction of flux in stromal Eno and PGlyM reactions. However, more experiments in this direction are required to take this any further.

Overall stoichiometries of the additional EMs of extended model formed three separate clusters, namely Clusters 8, 9 and 10, as shown in the dendrogram in Figure 4.11. Cluster 10, representing reactions participating in the EM 1, indicates a direct route through which redox equivalents in the stroma can be transferred to the mitochondria. Stromal NADPH is transferred into the cytosol via the stromal MAL/OAA shuttle, from where excess NADPH is moved into the mitochondrial matrix by the mitochondrial MAL/OAA shuttle. Inside mitochondria, however, NADH is converted to ATP by the activity of Complex I. Clusters 8 and 9 contain EMs that represent other routes through which reducing equivalents can be exported from the chloroplast to the cytosol. From the above observations it is evident that in plant cells under high light stress, these EMs mediate the conversion of excess stromal NADPH to mitochondrial ATP, thereby contributing to protecting the plant from photoinhibition (Section 2.2.4.4). Furthermore, along with EM 1, EMs in Clusters 8 and 9 represent the major routes through which chloroplast and mitochondria interact. From the direction of flow of flux in these EMs it is clear that the chloroplast act as the source and target for redox regulation in the plant cell.

From the heatmap in Figure 4.12 it is evident that G6P and PEP transporters participate in many EMs involved in the transfer of redox equivalents into the mitochondria. This result exemplifies the role of these transporters in mediating the transfer of energy and reducing equivalents into various compartments within the plant cell.

Part III

Integration

CHAPTER 5

Metabolic models to analyse microarray data

5.1 Introduction

Independent steady-state structural models of the light reactions, the Calvin cycle, glycolysis and the TCA cycle were described in Chapter 3. These models were then integrated in the previous chapter to construct a larger model where the characteristics of eukaryotic cells, such as the compartmentation of reactions and the involvement of transport reactions, were considered. This model was later used to investigate the ATP and redox interactions between chloroplasts, cytosol and mitochondria in a plant cell. The primary objective of this chapter is to use the model to analyse expression levels of genes obtained from microarray experiments. The basic principles and common techniques involved in the analysis and interpretation of microarray data were described in Section 1.5.

The motivation for performing this study was threefold. Firstly, previous attempts to integrate metabolic models with gene expression data have successfully employed ESs from steady-state stoichiometric models as the linking factor [38, 41]. One early result, that functionally related genes are often coexpressed [184], has provided strong motivation for the adoption of expression microarrays to study the characteristics of ESs. The outcome of these studies revealed important features of genetic and metabolic regulation existing within a metabolic model. A study by Schuster *et al.* [38], using a model of *S. cerevisiae* central carbon metabolism, showed that the expression levels of genes coding for enzymes in an ES are more highly correlated when compared to enzymes that were grouped randomly. Similar results were obtained from a recent analysis performed on a genome-scale model of *E. coli* [41]. With the help of the integrated expression values, the latter study also showed that many subsets in the *E. coli* model belong to known operons or regulons. These results supported the hypothesis that the enzymes within a given subset are genetically regulated in a coherent fashion. For these reasons, it was expected that combining stoichiometric model analysis techniques with gene expression profiles might reveal potential relationships and novel biological properties, and novel applications of metabolic models.

Secondly, unlike the case with ESs, no efforts have been made to integrate the RCCs of a steady-state stoichiometric model and the levels of gene expression. As described in Section 1.4.2.3, RCCs are a quantitative extension to the concept of ESs. They represent the strength of correlation between fluxes carried by reactions in a stoichiometric model, and are therefore a good candidate for relating with the correlation between the expression levels of genes. In reality, there exists a complex relationship between gene expression and metabolic flux. Many enzymes can accept several different substrates, thus relating the expression of one gene to several fluxes. The converse applies to enzyme complexes, where several genes are related to one flux. Similarly, in the case of isozymes, several genes are coupled to one or several fluxes. For these reasons, it is difficult to draw conclusions on flux through a metabolic reaction based on the expression of a corresponding gene and vice versa. However, integrating RCCs with microarray data may help to relate coexpressed genes and reactions that share similar flux.

Finally, integration of the properties of stoichiometric models such as RCCs may aid microarray data analysis by dimensional reduction, classification and annotation. Measurements of gene expression levels by microarray experiments create a high-throughput of data, the interpretation of which increasingly requires novel and efficient dimensional reduction strategies. Many clustering methods have been proposed and are widely used [185]. These algorithms group genes and/or samples into clusters of similar expression profiles, in order to suggest possible functional relationships between them [184, 186, 187, 188]. The importance of graphical representations and of clustering algorithms stands out from many recent publications devoted to co-expression analysis and gene function prediction [184, 186, 187, 188, 189, 190, 191].

This chapter aims to combine hierarchical clustering of RCCs representing the correlation between fluxes carried by reactions in a metabolic model, with the classical gene expression correlation based clustering for studying microarray expression data. Such an integration may provide novel graphical representations or cluster annotations through correlation between the fluxes carried by enzymes, and the expression levels of genes coding for these enzymes.

5.2 Methodology

5.2.1 Mapping the 'reaction—enzyme—protein—gene' associations in the model

In the definition of the models of the light reactions, the Calvin cycle, glycolysis and the TCA cycle given in Appendices A, B, C and D, respectively, the reaction names are represented using common reaction abbreviations. The foremost step in integrating these metabolic models with data from public databases is to substitute such abbreviations

with the reaction identifiers specific to the database of choice. The database used here was AraCyc (Section 1.5). A table containing the common reaction names used in the model and their corresponding AraCyc reaction identifiers is provided in Appendix E.

The next step is to identify all those genes that are coding for the reactions in the model. As has been described earlier in Section 1.5, the associations between reactions, proteins, enzymes and genes are not in most cases one-to-one, and hence not easy to map. A Python based ScrumPy add-on called PyoCyc† [13] was used for this purpose. It can be used to read BioCyc (Section 1.5) flat files[1] into the Python environment as a nested dictionary (Table 1.4). This dictionary has a hierarchical structure with genes as the root node[2] and proteins as children[3]. Enzymes are represented as children of proteins and reactions as the children of enzymes. Metabolites are the 'end-nodes' or 'leaves' of this structure. An example showing the Python scripts required for extracting genes coding for a reaction using PyoCyc is given below.

The mitochondrial reaction catalysed by the enzyme succinyl-CoA synthetase is represented in the model definition as 'SCS_mit'. This enzyme is known as 'SUCCCOASYN-RXN' in the AraCyc database. To extract the genes coding for this enzyme, first the AraCyc database has to be loaded into a Python dictionary object:

```
>>> import PyoCyc
>>> AraCyc = PyoCyc.Organism(data = "aracyc_database")
```

Once that is done, the dictionary has to be traversed to find the parents of this reaction, which will include enzyme-reaction associations, and the polypeptide monomers constituting the protein.

```
>>> AraCyc["SUCCCOASYN-RXN"].TravParents()
[ENZRXNQT-9439, AT5G23250-MONOMER, AT5G23250, ENZRXNQT-9438,
AT2G20420-MONOMER, AT2G20420, ENZRXNQT-9437, AT5G08300-MONOMER,
AT5G08300]
```

The parents of the items in the list above are the genes coding for 'SUCCCOASYN-RXN'.

```
>>> AraCyc["SUCCCOASYN-RXN"].TravParents()[0].GetParents()[0].
GetParents()[0].UID
'AT5G23250'
```

[1] A 'flat file' is a plain text or mixed text and binary file which usually contains one record per line, within which the single fields may be separated by delimiters, e.g. commas.

[2] Every hierarchical structure has a member that is at the highest level. This member is called the 'root' or root node. It can be thought of as the starting node.

[3] Nodes that share the same parent node.

Note that the script described above illustrates the extraction of only one gene. However, it can be modified to obtain all the genes coding for a particular reaction.

Genes coding for a few reactions, such as the transport reactions, could not be retrieved as the data pertaining to them was not available in the AraCyc database. Meanwhile, some other reactions in the model such as the 'Cyclic_lum' (mediating cyclic photosphosphorylation in the lumen) do not have any genes associated with them. The reason for this is that these reactions are either composite or spontaneous. In order to keep track of the isoforms of a protein for which a gene codes, the names of the isoforms were appended to the gene name as suffix.

5.2.2 Integrating metabolic models with gene expression data

The Nottingham *Arabidopsis* Stock Centre's microarray database, NASCArrays (Section 1.5), was the source of gene expression data used in this study. The 'super bulk gene' file containing nearly 3500 hybridisations, each with expression levels of over 22,500 genes represented on the ATH1 array, was downloaded. These arrays were derived from varied experiments, tissues, conditions, treatments and genetic backgrounds, providing the diversity for correlation analysis. For the purpose of this study, however, only those arrays derived from experiments conducted on leaves, rosettes and cotyledons were used. From the 63 arrays obtained, three that used RNA from species other than *A. thaliana*, or which involved pre-amplification of the RNA used as the source for the hybridisation, were excluded. Expression data from individual experiments were log-transformed to adjust for the effects of variations in the quantity of starting RNA and the differences in the labelling and detection efficiencies. No further modification or scaling was made on the data unless otherwise specified. An expression matrix representing all experiments used in this analysis along with the genes, a short description of the genes and their expression values can be found in the tab delimited[4] text file 'supercluster.txt' in the compact disc (CD) accompanying this thesis.

Expression data for the genes ultimately coding for reactions in the model were extracted and large-scale correlation analysis was performed essentially as described by Causton *et al.* (2003) [185], by calculating Pearson's correlation coefficient for each gene pair (Section 1.5).

5.2.3 Clustering of the correlation matrix and generation of compressed heatmaps

The expression profiles of genes in the Pearson's correlation matrix generated in the previous section were hierarchically clustered using the WPGMA algorithm (Sec-

[4] A tab delimited file can be imported into standard spreadsheet programs such as Microsoft Excel and OpenOffice.org Spreadsheet.

tion 1.4.2.3), and an expression correlation tree was generated as described in Section 1.4.2.3. Leaves of this tree represent genes in the model and the intermediate nodes are clusters that represent genes that are coexpressed. The columns of the correlation matrix were then sorted in the order of the leaves of the expression correlation tree.

A reaction correlation tree based on RCCs was generated from the model using the method described in Section 1.4.2.3. The order of the reactions in this dendrogram was used to sort the corresponding genes along the rows of the expression correlation matrix.

The correlation matrix obtained after clustering the rows and columns was then imported into the TM4-MeV† [44] multi experiment visualisation program (version 4.5.1) to generate a compressed gene expression correlation heatmap. Although, MeV provides a number of clustering algorithms and methods for microarray data analysis, note that it was used here solely as a heatmap visualisation tool.

The annotated Python code of the program used for the automation of the steps involved in the construction of clustered Pearson's correlation matrix from a gene expression matrix, using input from a metabolic model, is included in the CD. A UML†[5] diagram representing the interaction between the Python classes in this program is shown in Appendix F.

5.3 Results and Discussion

The Python based ScrumPy tool PyoCyc was effectively used to extract the possibly complete set of genes coding for reactions in the combined model containing the light reactions, the Calvin cycle, glycolysis and the TCA cycle. The final set of 54 reactions yielded a total of 193 genes for further analysis. The gene expression profiles of these genes from 60 varied microarray experiments performed on green leaves, rosettes or cotyledons were then extracted from NASCArrays. The gene expression matrix obtained had genes coding for reactions in the model along the rows, and various microarray experiments along the columns.

A Pearson's correlation matrix representing the correlations between the expression profiles of genes coding for reactions in the model was constructed. It was visualised using the TM4-MeV heatmap visualisation tool and the compressed heatmap obtained is shown in Figure 5.1. Although no detailed information was obtained from this heatmap, hierarchically clustering the rows and columns of the matrix grouped together genes whose expression profiles correlate (Figure 5.1). A dendrogram showing the four distinct clusters in this heatmap is given in Figure 5.2. Each of these clusters represents genes sharing correlated expression profiles. Numerous studies [184, 44, 186, 187, 188, 190] have shown that genes in such clusters are functionally related and are often

[5] The Unified Modeling Language is an open method used to specify, visualise, construct and document the artifacts of an object-oriented software

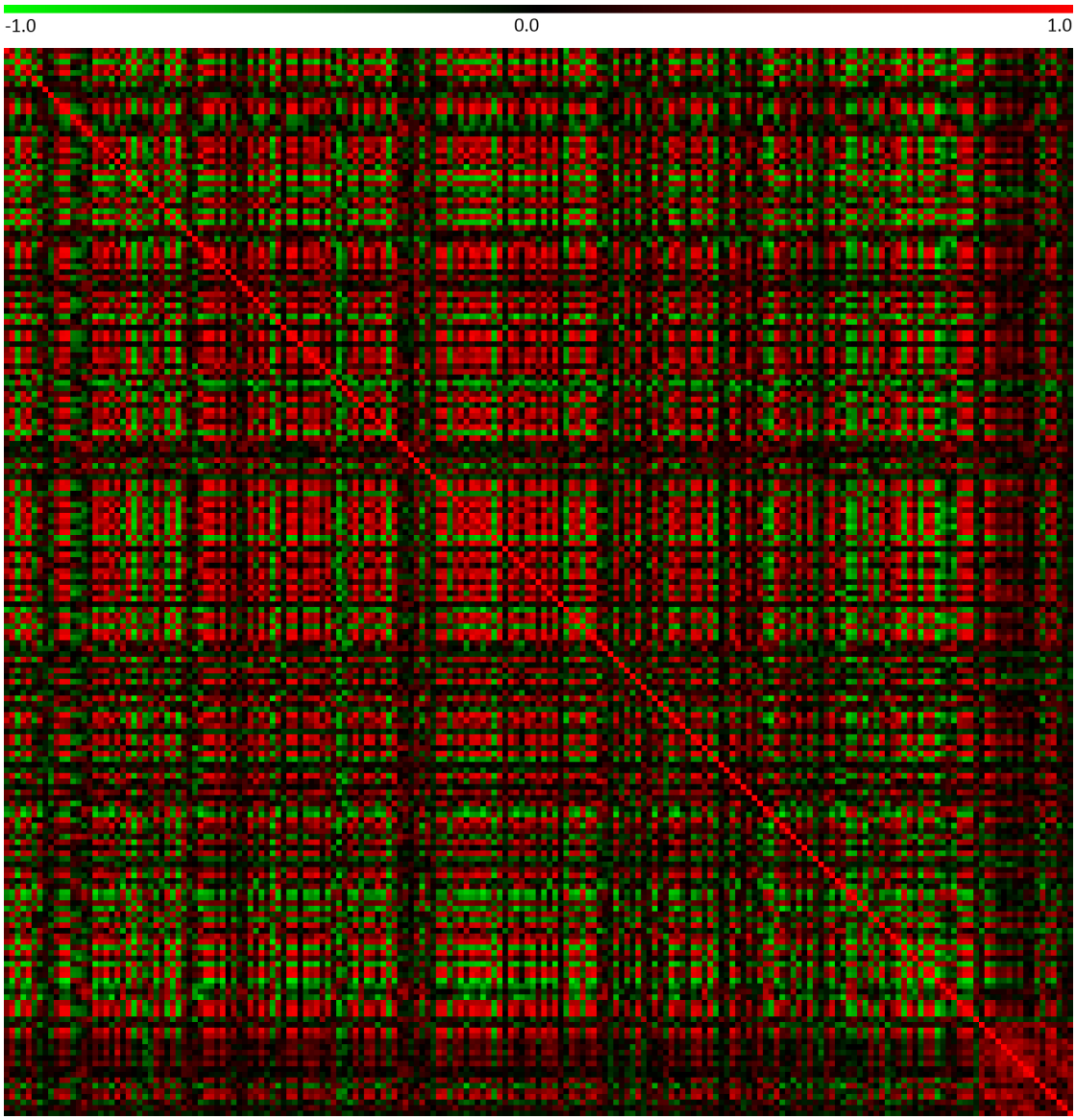


Figure 5.1 – Heatmap showing Pearson's correlation matrix generated from the expression matrix representing the correlations between genes coding for reactions in the combined model. Positive and negative correlation between the genes are represented by red and green, respectively. Black represents no correlation between genes. Red cells across the diagonal of the matrix shows individual genes perfectly correlating with themselves. Row and column designations are omitted here for clarity, but are included in the version provided in the CD.

coexpressed. Similar properties were observed in the Pearson's correlation matrix in Figure 5.3. Cluster A contained genes that primarily code for reactions involved in photosynthetic light reactions. Several of the Calvin cycle and TCA cycle genes were also observed in this cluster. Clusters B and C contained a mix of genes that code for the Calvin cycle and glycolysis, and a few reactions of the TCA cycle. Cluster D, on the other hand, had genes that predominantly code for reactions of the TCA cycle. The distribution of genes in these clusters provided enough evidence to indicate that functionally related genes in the model are coexpressed. A couple of examples

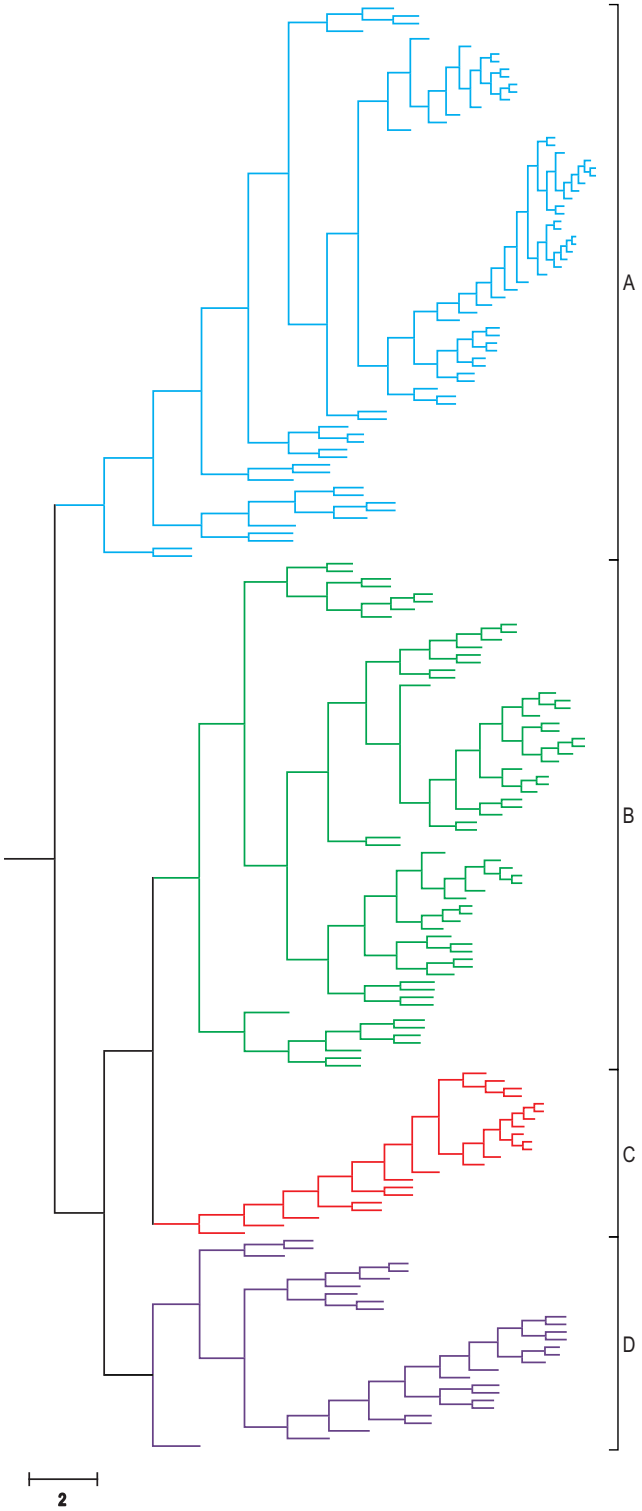


Figure 5.2 – Dendrogram representing the correlations between the expression profiles of genes coding for reactions in the model. Leaf names are not shown here for clarity, but are provided in the file 'chlorocytomito_exprFullTree.pdf' included in the CD. Four separate clusters were observed. Cluster A contains genes that predominantly code for reactions of the photosynthetic light reactions. However, many genes coding for the Calvin cycle and glycolysis reactions were also observed. Clusters B and C contain a mix of genes that code for reactions of the Calvin cycle and glycolysis, and very few that code for the TCA cycle reactions. Cluster D has genes that predominantly code for reactions of the TCA cycle. Scale bar respresents a difference of $\theta^{K}_{xy} = 2$rad.

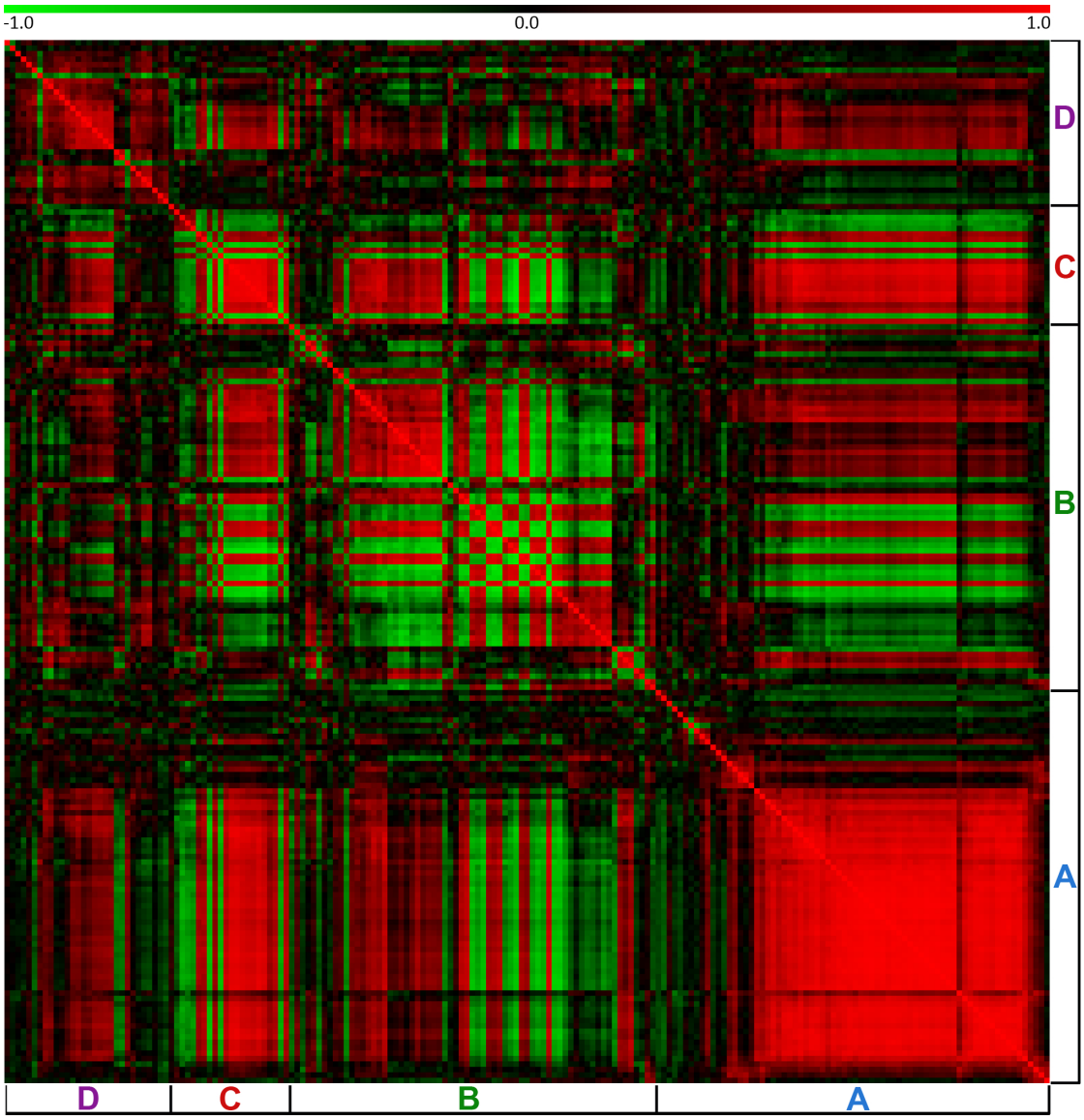


Figure 5.3 – Heatmap showing Pearson's correlation matrix representing the correlations between the expression profiles of genes coding for reactions in the combined model. Rows and columns of this matrix were hierarchically clustered. The dendrogram based on which the clustering was performed is shown in Figure 5.2. See the text associated with this dendrogram for details on the four clusters A, B, C and D. Positive and negative correlation between the genes are represented by red and green, respectively. Black represents no correlation between genes. Red cells across the diagonal of the matrix shows genes perfectly correlating with each other. Row and column designations are omitted here for clarity, but are included in the version provided in the CD.

of such genes are those coding for the reactions 'RXN-924' of the light reactions (Cluster A) and 'SUCCINATE-DEHYDROGENASE-(UBIQUINONE)-RXN' of the TCA cycle (Cluster D). A few earlier studies have shown that coexpressed genes in clusters occupy nonrandom positions with respect to the pathway structure [186, 188]. However, no affiliation to any specific pathway structure was immediately evident from the distribution of genes in the correlation matrix.

A metabolic tree was constructed based on the RCCs calculated from the null space of the stoichiometry matrix of the combined model of the light reactions, the Calvin

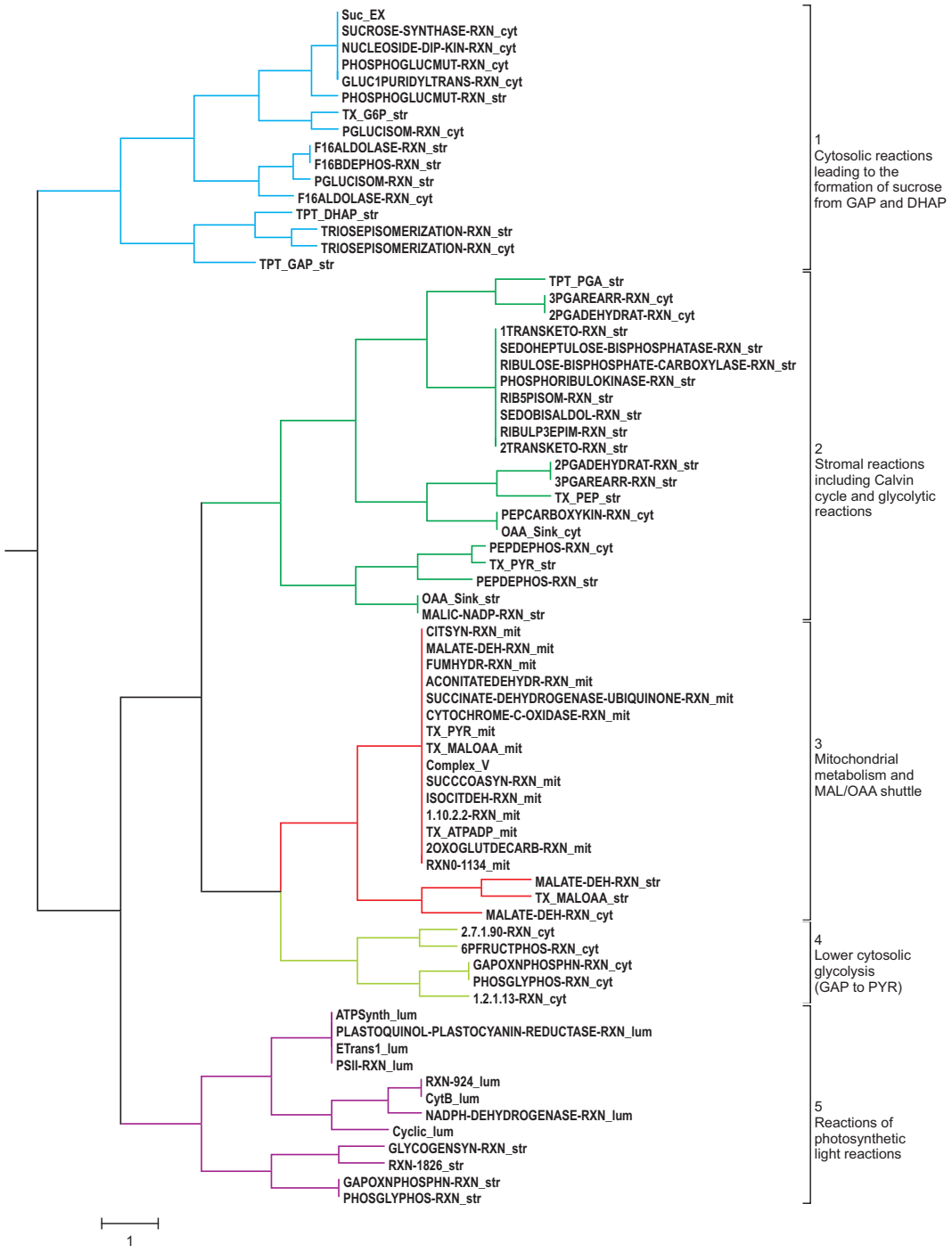


Figure 5.4 – Dendrogram representing the correlations between fluxes carried by reactions in the combined model of light reactions, the Calvin cycle, glycolysis and the TCA cycle in the presence of carbon flux. Five distinct clusters were observed; the general characteristics of each are presented. '_lum', '_str', '_cyt' and '_mit' indicate localisation of metabolites in the lumen, stroma, cytosol and mitochondria, respectively. Reaction designations were obtained from AraCyc 6.0. See Appendix E for reaction name definitions. Scale bar respresents a difference of θ^{K}_{xy} = 1rad.

cycle, glycolysis and the TCA cycle in the presence of carbon flux (Figure 5.4). Clusters on this tree represent correlations between fluxes carried by reactions in the model. Five functional clusters were observed. Cluster 1 represents the correlations between

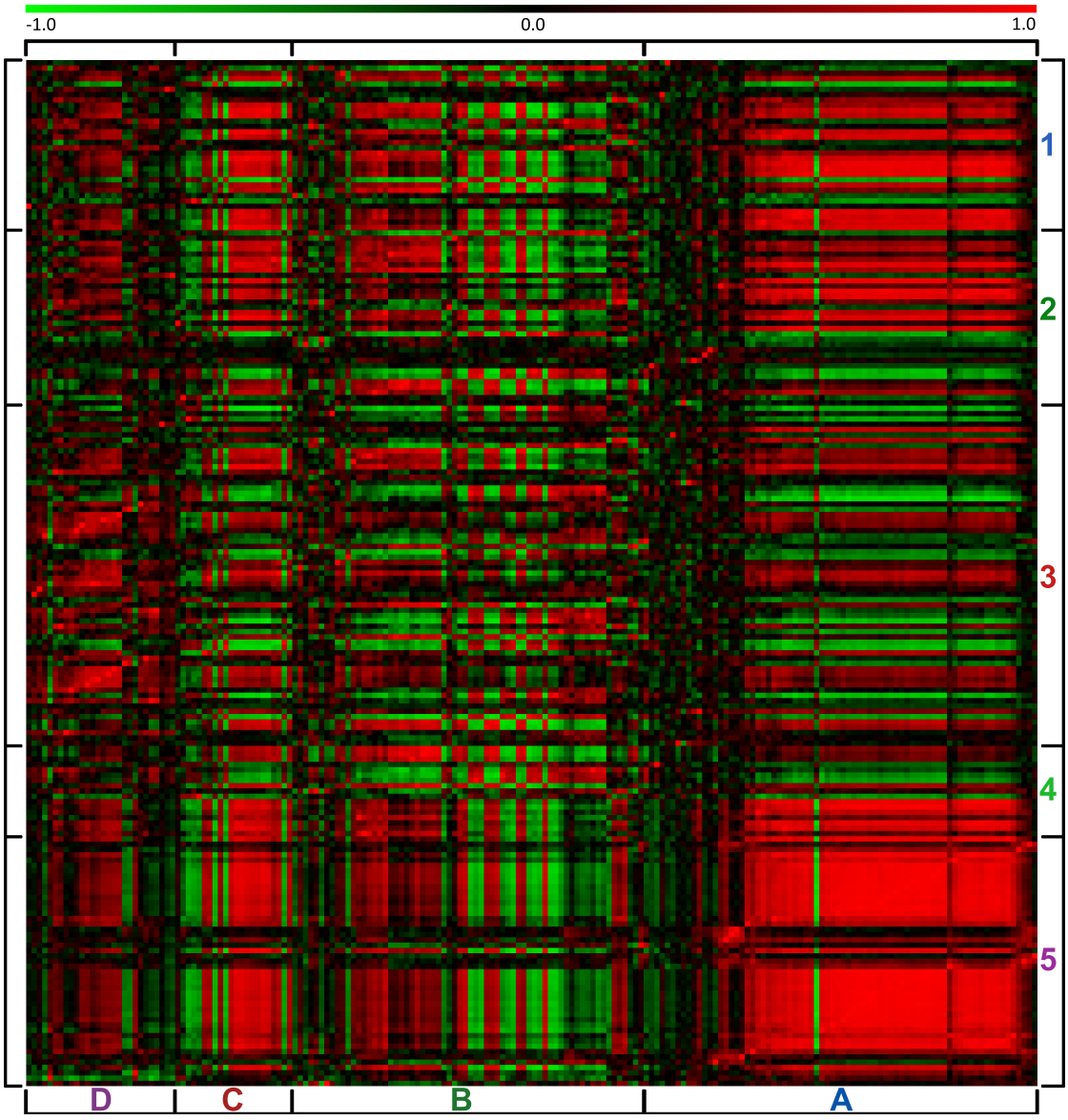


Figure 5.5 – Heatmap showing the Pearson's correlation matrix, whose rows and columns were hierarchically clustered based on reaction correlation coefficients (RCCs) representing correlations between fluxes carried by reactions in the model and Pearson's correlation coefficients representing correlations between the expression profiles of genes coding for reactions in the model, respectively. The dendrogram based on which the clustering was performed is shown in Figures 5.4 and 5.2, respectively. See the text assosiated with these dendrograms for details of the various clusters shown here. Positive and negative correlation between the genes are represented by red and green, respectively. Black represents no correlation between genes. Row and column designations are omitted here for clarity, but are included in the version provided in the CD.

reactions of upper glycolysis/gluconeogenesis, mediating the conversion of chloroplast intermediates to sucrose, whereas Cluster 4 contains reactions of lower glycolysis, converting GAP and DHAP to PYR. Cluster 2 groups reactions of the Calvin cycle and the glycolytic reactions in the stroma. Correlations between reactions of the TCA cycle are represented in Cluster 3. Finally, Cluster 5 indicates correlations between reactions of the photosynthetic light reactions. These clusters point to the distribution

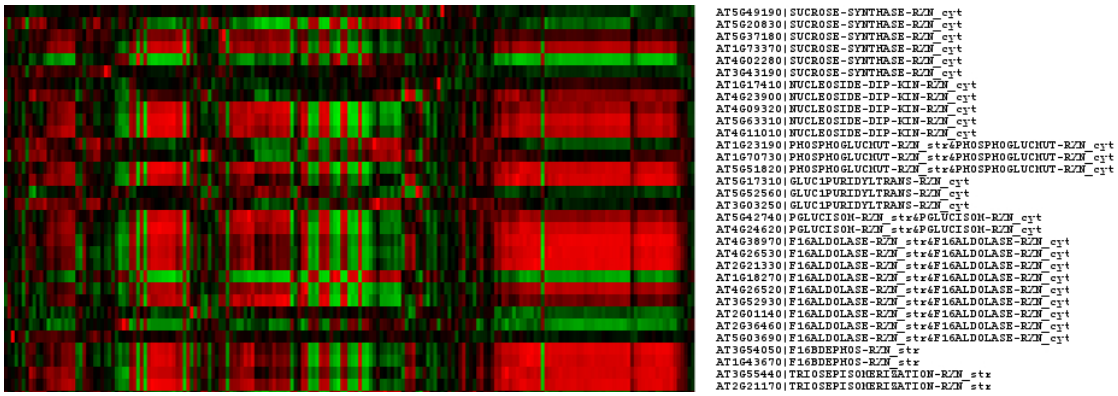


Figure 5.6 – Corresponding correlation profiles of the genes coding for reactions in Cluster 1 that have correlated flux. Gene names are suffixed with reaction names and details about the compartment in the metabolic model from which it was extracted. '_str' and '_cyt' represent stroma and cytosol, respectively. Red and green represent positive and negative correlation, respectively. Black indicates 'no correlation'. See Appendix E for reaction name definitions.

of reactions in the various compartments included in the model. The order of the reactions along the leaves of this tree was then used to sort the genes along the rows of the Pearson's correlation matrix constructed earlier, in such a way that the genes coding for a particular reaction are placed together. The resulting heatmap, shown in Figure 5.5, now contains genes whose expression profiles correlate, and are hence considered coexpressed, arranged along the columns, and genes that code for reactions that have correlated fluxes clustered arranged the rows. Note that, from here on, each cluster on this matrix will be referred to using co-ordinates; for example, Cluster A5 represents the cluster formed at the intersection of the correlation profiles of genes in Cluster A across the column and Cluster 5 across the row.

Several observations were made from the pattern of distribution of the reordered clusters on this heatmap. Firstly, it was found that many genes coding for reactions that have correlated metabolic fluxes often have corresponding correlation profiles. This characteristic is particularly evident among the positively correlated genes that code for reactions in each of the horizontal clusters. A representative example, Cluster 1, with gene and reaction designations is shown in Figure 5.6. A number of genes coding for each reaction in the cluster have similar correlation profiles with many genes coding for other reactions in the cluster. Similar observations can be made from the correlation profiles of genes coding for reactions in Clusters 2 and 5 in Figure 5.7. This result has strong implications as it demonstrates that genes coding for reactions that share similar fluxes are often coexpressed.

Secondly, the clusters in the heatmap can now distinguish within- and cross-pathway correlation patterns. As described earlier, each metabolic pathway is composed of a series of biochemical reactions that are connected by their intermediates: the reactants (or substrates) of one reaction are the products of the previous one, and so on. For this

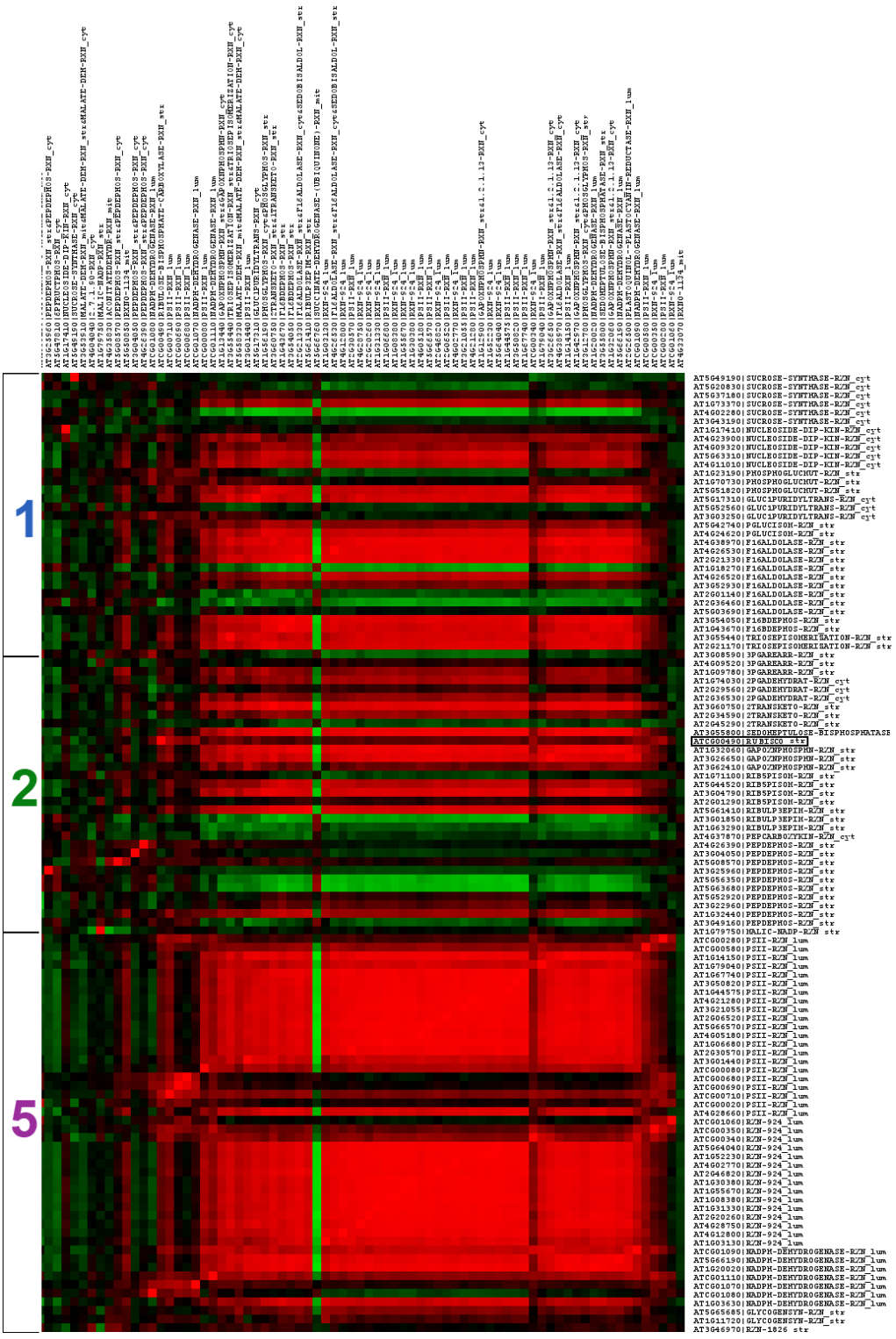


Figure 5.7 – Heatmap representing within- and cross- pathway correlations. It also shows the correspondence between the correlation profiles of rubisco (highlighted with with a box around the gene name) and other genes coding for reactions in chloroplast and cytosol. Note that the Clusters B, C and D were trimmed to enable this visualisation, but are considered as part of the correlation profile. Gene names are suffixed with reaction names and details about the compartment in the metabolic model from which it was extracted. '_str' and '_cyt' represent stroma and cytosol, respectively. Some gene name suffix were either shortened or trimmed off for clarity. Red and green represent positive and negative correlation, respectively. Black indicates 'no correlation'. See Appendix E for reaction name definitions.

reason, there is a good chance for such reactions within a pathway to be coexpressed. Similarly, many pathways are interconnected with other pathways by means of utilising intermediate metabolites or end products from other pathways. Therefore, it is likely that reactions in some pathways have similar co-expression profiles with genes coding for reactions in other pathways. The heatmap shown in Figure 5.6, representing the genes coding for reactions involved in the conversion of cytosolic GAP and DHAP to sucrose (Cluster 1), exemplifies a within-pathway correlation pattern. Figure 5.7, on the other hand, shows a heatmap visualisation in which Clusters 5, 2 and 1 — representing reactions involved in the photosynthetic light reactions, the Calvin cycle and upper glycolysis (GAP to sucrose) — share corresponding correlation profiles.

Thirdly, the heatmap is now able to distinguish the compartments in which a gene is more highly expressed. From the dendrogram shown in Figure 5.2, it is evident that each cluster contained a high number of genes coding for reactions in a particular compartment. For example, Clusters A and D contained genes predominantly coding for light reactions and the Calvin cycle in the chloroplast, and the TCA cycle reactions in the mitochondria, respectively. Similarly, the metabolic tree shown in Figure 5.4 was able to successfully distinguish the various compartments in the model, based on the correlation between the fluxes carried by the reactions within them. By clustering the rows and columns of the Pearson's correlation matrix based on these dendrograms, distinct patterns were formed on the heatmap that could reveal the localisation of enzymes. For instance, Cluster D3 represents genes that strongly correlate with genes that code for reactions in the mitochondria. Another example is Clusters A1, A2 and A5, where the genes strongly correlate with genes that code for stromal and cytosolic reactions (Figure 5.7). Accordingly, it was found that correlation profiles of genes in these clusters correspond to the correlation profiles of genes with known localisation (Table 5.1). An example is the gene 'ATCG00490' coding for rubisco in the chloroplast (Figure 5.7). Note that the correlation profiles of this gene correspond to those coding for light reactions in Cluster 5 and other reactions in the chloroplast (Clusters 1 and 2). It follows from this observation that by comparing the correlation profile of a gene with known localisation with the correlation profiles of other genes in the cluster, conclusions can be drawn about their compartmentation in the cell. Based on this hypothesis, localisation of the complete set of proteins coded by the genes in the matrix can be predicted using standard programming techniques. Such programs have the added benefit of allowing the user to control the sensitivity of the prediction by selecting a threshold correlation coefficient based on which corresponding correlation profiles can be identified. For example, a threshold correlation coefficient of 0.8 will only identify genes that strongly correlate with the reference gene whose localisation is already known.

Identifying the localisation of proteins is an important step towards a broader understanding of the cellular function as a whole, and may help in determining the

Table 5.1 – A random list of genes coding for reactions in the extended model, and the localisation predictions made by various bioinformatic and experimental techniques. 'chl', 'cyt', 'mit', 'per', 'nuc', 'ext' and 'unc' represent chloroplast, cytosol, mitochondria, peroxisome, nucleus, extra cellular and unclear, respectively. '-' means that the software was unable to make any prediction.

	TargetP	MitoProt2	SubLoc	IPSort	Predotar	MitoPred	Peroxp	WolfPSort	MultiLoc	LocTree	MassSpec	GFP	Microarray
AT2G21330	chl	mit	mit	mit	chl	-	-	chl	chl	chl	chl	chl	chl
AT3G04790	chl	mit	cyt	chl	chl	mit	-	mit	chl	chl	chl	-	chl
AT4G12800	chl	mit	ext	mit	chl	mit	-	chl	chl	chl	chl	-	chl
AT1G52230	chl	mit	mit	chl	chl	mit		chl	chl	chl	chl		chl
AT3G12780	chl	mit	cyt	mit	chl	-	-	chl	chl	chl	mit	cyt	chl
AT1G18270	ext	-	cyt	chl	er	-	-	per	chl	cyt	nuc	-	cyt
AT5G52920	chl	mit	mit	mit	chl	-	-	chl	chl	chl	chl	chl	cyt
AT2G22480	chl	mit	cyt	chl	chl	mit	-	chl	chl	chl	-	-	cyt
AT1G12000	chl	-	cyt	mit	chl	-	-	chl	-	cyt	-	-	cyt
AT3G04120	-	mit	cyt	-	-	-	per	per	-	cyt	mit	cyt	cyt
AT2G20360	mit	mit	mit	mit	mit	mit	-	chl	mit	chl	mit	-	mit
AT5G14590	chl	mit	mit	mit	mit	-	-	chl	mit	chl	mit	-	mit
AT2G27730	mit	mit	mit	-	mit	mit	-	chl	mit	-	mit	mit	mit
AT1G47420	mit	-	mit	mit	mit	mit	-	mit	mit	cyt	mit	-	mit
AT2G05710	chl	mit	cyt	chl	chl	-	-	unc	chl	chl	mit	-	mit

role of thousands of uncharacterised proteins predicted by the genome sequencing projects. Modern organelle-focused experimental approaches can identify proteins in a given compartment. However, reliable protein localisation requires that the technique used must be able to distinguish between genuine organelle residents and contaminating proteins [192]. Although reasonably pure preparations of some organelles can be achieved, there are many difficulties associated with measuring and characterising proteins that are in a compartment [193]. Nevertheless, a variety of experimental methods are currently being used to identify protein localisation. Recently green fluorescent protein (GFP) and mass spectrometry (MS) techniques have been successfully employed to deduce the localisation of approximately 1100 and 2600 proteins, respectively [194, 193, 195]. Although these techniques have accelerated the flow of protein localisation information, the subcellular location of the majority of proteins in a plant cell is still not known.

A relatively simple, low-cost and rapid means to tackle this issue is to employ bioinformatic targeting algorithms to predict protein localisation from amino acid sequence. A number of software tools exist, including TargetP† [196], Predotar† [197], iPSORT† [198], SubLoc† [199], MitoProt II† [200], MITOPRED† [201], PeroxiP† [202], and WoLF PSORT† [203], which can predict proteins targeted towards plastid, cytosol, nucleus, mitochondria, peroxisome or the endoplasmic reticulum. However, many of these tools are aimed at identifying particular compartments, and hence predictions

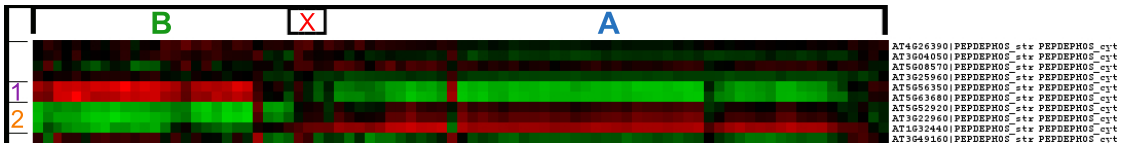


Figure 5.8 – Heatmap representing the correlation profiles of genes coding for isoforms of pyruvate kinase (PK) in different compartments. Gene names are suffixed with the AraCyc reaction identifier and details about the compartment in the metabolic model from which it was extracted. Some clusters in the region 'X' were trimmed off to enable this visualisation. Clusters A contains genes that predominantly code for reactions in the chloroplast. Cluster A1 represents PK genes anticorrelating with other genes in Cluster A. On the other hand, Cluster A2 contain genes that strongly correlate with Cluster A. Clusters B1 and B2 represent similar characteristics of correlation with genes that predominantly code for reactions in the cytosol. '_str' and '_cyt' represent stroma and cytosol, respectively. Red and green represent positive and negative correlation, respectively. Black indicate 'no correlation'.

made by them are reportedly biased towards that compartment [204]. Furthermore, the outputs of such programs have been found to be somewhat inconsistent with each other, or with experimentally determined results [204], making them unreliable for some analyses. Nevertheless, localisation predictions made by these tools on the complete set of genes in *A. thaliana* are available in the SUBA II[†] database [195] for free download. The approach described earlier was used to predict the localisation of the complete set of genes in the model and the results were compared against the predictions made by experimental techniques such as GFP and MS as obtained from the SUBA database. A random list of genes in the model and their predicted localisation along with the available bioinformatic and experimental predictions are shown in Table 5.1. The 'Microarray' column in this table contains the predictions made using the approach described in this thesis. It is evident that the bioinformatic tools MitoProt2 and MitoPred are strongly biased towards mitochondrial predictions. On the other hand, TargetP and MultiLoc seem to show bias towards identifying genes in chloroplasts. It was found that the predictions made by the method described in this study strongly correspond to those obtained from GFP and MS based predictions. A similar table containing a list of 107 genes coding for reactions in the extended model and their localisation predictions is included in the accompanying CD as a tab delimited text file named 'predictions.txt'. Localisation of the remaining genes could not be predicted as they did not correlate strongly with the reference genes used. A threshold correlation coefficient of 0.7 was used to obtain this result.

Finally, the heatmap made it possible to distinguish genes coding for isoforms of proteins in different compartments. It was found that correlation profiles of some genes coding for a particular reaction do not correspond to the correlation profiles shared by the other genes coding for the same reaction. Correlation profiles of such genes, however, were found to correspond with those in different compartments. This observation suggests that these two sets of genes may be coding for isoforms of proteins

in two different compartments. For instance, the correlation profiles of genes coding for 'PEPDEPHOS-RXN' (pyruvate kinase) are shown in the heatmap in Figure 5.8. Three genes that share similar correlation profiles can be seen (Cluster 2A) corresponding strongly with genes coding for reactions in the chloroplast. Meanwhile, two other genes with similar properties can be seen (Cluster 1B) corresponding strongly with genes coding for cytosolic reactions. This difference in their correlation profiles can mean that they are coding for protein isoforms localised in chloroplast and cytosol. Similar cases involving genes coding for the reactions 'GAPOXNPHOSPHN-RXN', 'PHOSGLYPHOS-RXN', 'MALATE-DEH-RXN' and 'PHOSPHOGLUCMUT-RXN' were detected in the heatmap, but discussing them all would just reiterate the same point.

5.4 Conclusion

Results obtained from this analysis shows that the characteristics of metabolism derived from metabolic models can be effectively integrated with large-scale experimental data. The approach described here was able to successfully relate the correlation between fluxes carried by reactions in a stoichiometric model with the expression profiles of the genes involved. Furthermore, the results suggest that metabolic models can be used to upgrade the information content in gene expression data by providing additional metabolic information. The case study involving a compartmentalised stoichiometric model of plant metabolism and related gene expression data provided new insight into within- and cross-pathway correlation patterns. This result has strong implications as it shows that large-scale metabolic models can be used to identify new genes and can assist in gene annotation. Another outcome of this case study was in reference to the localisation of proteins in particular compartments within a plant cell. The results showed that the correlation profiles of a gene can aid in predicting the localisation of its protein product. Given good quality microarray expression data containing sufficient experiments that allow reliable statistical analysis, the technique described here can be used more generically. With the large number of publically available metabolic networks and expression data, this approach may significantly contribute to the identification of enzyme localisation in many different eukaryotic systems. Finally, the case study showed that the correlation profiles, together with RCCs, can be used to distinguish genes coding for protein isoforms. Such information may again help in gene annotation.

Part IV

Discussion

CHAPTER 6

General discussion and future directions

6.1 Relevance and implications of the modelling described in this thesis

The initial objective of this thesis was to construct a stoichiometric model of the central carbon metabolism in a plant cell. This objective seemed very trivial considering the relatively simple data requirements for defining stoichiometric models and the numerous successful attempts at constructing stoichiometric metabolic models, although mostly prokaryotic. One challenge that became immediately apparent in the initial stages of this study was associated with the localisation of enzymes and metabolites in specific compartments within the plant cell. Compartments act by sequestering the enzymes and metabolites participating in specific metabolic processes and thereby preventing the simultaneous occurrence of potentially incompatible reactions elsewhere within the cell (Section 2.2). Therefore, any attempt to construct a stoichiometric model of eukaryotic system, in particular that of higher eukaryotes such as plants and animals, is not complete if this property has not been addressed.

Unlike modelling metabolism within a single major compartment, as in the case of prokaryotes[1], modelling metabolism in multi-compartmental systems requires careful mapping of the numerous interactions between the compartments sequestering the metabolic pathways of interest. Furthermore, it is also essential to devise a strategy to integrate the localisation information in the model definition. For these reasons, a modular approach was adopted in the modelling described in this thesis. Self-contained steady-state stoichiometric models of metabolism in specific compartments were constructed and their independent properties and behaviours were analysed using a selection of standard model analysis techniques that were described in Chapter 1. Metabolites and reactions in these models were segregated using unique name suffixes that distinguish their localisation. The considerably smaller independent models were then integrated using specific transport reactions to form a larger model that simultaneously maintains

[1] A number of prokaryotic models include a 'periplasm' and/or a 'cell wall' compartment — though not so much happens in these of relevance to central carbon metabolism.

the localisation of reactions and metabolites and the interactions between the various compartments involved.

While a number of organism-specific genome-scale stoichiometric metabolic models of prokaryotes and lower eukaryotes have been constructed in the post-genomic era (approximately 15 in 2006 [205]), very few such models of higher eukaryotes have been published [13, 206]. These models, by not taking the localisation of metabolites and reactions into consideration, exemplify the major challenge that has been restricting researchers from attempting to construct genome-scale models of eukaryotic systems. The approach described in this thesis is simple and effective and can be extended to define compartmentalised genome-scale models.

6.2 Model analysis techniques to analyse compartmentalised models

The purpose of Chapter 1 was not only to survey the multitude of techniques involved in stoichiometric model analysis, but also to highlight how such methods can be used for analysing complex models of eukaryotic systems. Over the past several years EM analysis has been extensively used to investigate the characteristics of steady-state stoichiometric models of prokaryote metabolism (Section 1.4.3.1). It has found numerous applications, especially in identifying novel pathways within metabolic models and predicting the biomass yield (Section 1.4.3.1). This thesis describes the application of EM analysis to study the interaction between compartments in a eukaryotic system. In Chapter 4 EM analysis was successfully employed as a tool to identify major routes involved in the transfer of ATP and reducing equivalents between multiple compartments within the plant cell. The results obtained from this study illustrate the application of EM analysis in predicting the role of specific transporters in the exchange of ATP and reducing equivalents. Furthermore, it describes how EM analysis can be fruitfully employed to investigate the role of transporters in integrating the metabolism in various compartments.

Chapters 3 and 4 demonstrate novel techniques for the visualisation of the overall stoichiometries of EMs and the reactions participating in the complete set of EMs of a system. These techniques not only enable clustering, and thereby assimilating the information content in large number of EMs, but also help in extracting valuable biological information. With the advent of new and improved algorithms for the identification of the complete set of EMs in a genome-scale model [183], the modelling community is on the lookout for techniques that can extract valuable information from the millions of EMs that could be generated. The heatmap visualisation technique employed in this thesis to analyse reactions participating in EMs is a proven candidate for analysing such high-throughput data.

RCCs and the metabolic trees constructed from them were extensively used throughout this thesis. They were instrumental in identifying reactions that shared correlated fluxes within specific models described in Chapters 3 and 4. Using RCCs to study metabolic models is a very recent and underused concept, and the results illustrated in this study can serve as additional examples for its ability to identify functional modules within metabolic models. Above all Chapter 5 of this thesis describes a novel application of RCCs — as means for integrating metabolic models with gene expression data obtained from microarray experiments. A number of previous studies have attempted to use ESs to integrate properties of metabolic models with gene expression data [38, 41]. Such studies do not take into consideration either the relationship between ESs in a model or the relationship between reactions in ESs and those that are not. These studies, therefore, often neglect genes coding for reactions in the model that are not involved in a subset. During the course of this study it was found that the above disadvantages can be overcome by using RCCs to integrate model properties and gene expression data. The range of applications of RCCs described in this study advocates its wide use as a standard stoichiometric model analysis tool.

6.3 Metabolic models, microarray data and localisation predictions

Microarray experiments are high-throughput techniques that generate expression profiles of the complete set of genes in an organism. Traditional microarray data analysis techniques involve the application of various clustering algorithms to study the correlation between expression profile of genes. While these techniques often neglect the vast amount of biochemical information available, this thesis wanted to emphasise its importance in obtaining improved predictions of cellular behaviour. The approach described in Chapter 5 used the correlation between fluxes carried by reactions in a steady-state stoichiometric model to cluster and analyse expression profiles of genes coding for reactions in the model. The results obtained suggest that biochemical properties of metabolic models can be used to upgrade the information content in gene expression data by providing additional metabolic information. The outcome of this study has strong implications as it demonstrates a potential application of eukaryotic genome-scale metabolic models. Incorporating such models into microarray data may facilitate gene annotation and the identification of genes coding for protein isoforms. Furthermore, it can aid in predicting the localisation of enzymes in multi-compartmental systems. At present, there are very few methods that can effectively predict the localisation of enzymes in a eukaryotic cell. While some of the most reliable methods are based on molecular biology techniques and hence require considerable time and

effort in obtaining data, others are based on bioinformatic algorithms whose results were often found to be biased [204]. However, in the presence of a steady-state genome-scale model and sufficient microarray data, the technique described in this thesis can be used as a quick and preliminary means of determining the localisation of the genes of interest.

6.4 Directions for future work

ATP and NADPH generated during the light reactions are exported into the cytosol via various shuttle mechanisms in the chloroplast membrane. Once in the cytosol, NADH has several important uses. It is taken up by mitochondria to produce ATP and by peroxisomes for hydroxypyruvate reduction occurring as part of the photosrespiratory pathway [151]. It is also essential for nitrate reduction proceeding as a partial step of nitrate assimilation in the cytosol [151, 119]. Analysis of the integrated model of the light reactions, the Calvin cycle, glycolysis and the TCA cycle described in this thesis illustrated the various shuttle mechanisms involved in the transfer of ATP and reducing equivalents into the cytosol and identified the important transport proteins involved. Furthermore, it described some of the routes involved in the uptake of cytosolic reducing equivalents into the mitochondria. The integrated model, however, does not contain reactions of the photorespiratory pathway or the nitrate reduction pathway. Therefore, an important direction for future work is to extend the integrated model with reactions in these pathways. Analysis of the extended model may reveal novel routes involved in the exchange of ATP and reducing equivalents between chloroplast and mitochondria, and may provide more insight into the important aspects of ATP and NADH utilisation in the cytosol. Above all, it will enable a more comprehensive analysis of the energy and redox interactions between various compartments within the plant cell.

Integrating reactions of the nitrate reduction pathway into the model has an added advantage as it will then enable the introduction of new reactions into the model such as those involved in nitrogen assimilation. Nitrogen assimilation takes place inside the chloroplast and consumes ATP generated during light reactions [151]. Therefore, integrating them would help to draw a more complete picture of the energy utilisation within chloroplast. Integrating nitrogen assimilation reactions will also introduce more transport reactions into the model such as those involved in the MAL/CIT and the MAL/ASP shuttle mechanisms on the mitochondrial membrane. Similarly, transport reactions of the chloroplast membrane such as the MAL/2-OG shuttle will become included in the model. These shuttle mechanisms may reveal more routes involved in the exchange of energy and reducing equivalents between stroma and mitochondrial matrix. Furthermore, they will help to investigate the numerous interactions between the cytosol and mitochondria.

In the same way, the integrated model can be extended by including other metabolic pathways in various compartments such as the shikimate pathway in the stroma. This

will contribute towards achieving a compartmentalised genome-scale model of the plant cell. Note that the automated reconstruction of compartmentalised eukaryotic genome-scale models is as yet impossible because of two important factors. Firstly, information about the localisation of the majority of enzymes and metabolites in a plant cell is not yet available. Secondly, the stoichiometric information pertaining to transport reactions is not clearly defined in pathway databases such as KEGG and AraCyc that are used for the automated reconstruction of genome-scale models.

Furthermore, such extended models, once integrated with experimental data such as gene expression profiles, may help in identifying novel biological properties of the system. For example, it can reveal novel isoforms of genes in various compartments and thereby help in gene annotation. Most importantly, such extended models can aid in testing the approach described in Chapter 5 to identify major issues that have to be addressed, and in refining the technique further.

Another important direction for future work involves analysing the extended model in the absence of the light reactions. Plants use the ATP and NADPH produced during the light reactions to produce transitory starch. In the absence of light, however, it is broken down by various enzymes in the chloroplast such as amylases and phosphorylases to produce glucose, maltose and other intermediates. The former is oxidised by reactions of the OPPP to produce redox potential in the form of NADPH and substrates required for initiating shikimate pathway and nucleotide synthesis [125, 63]. Most intermediates of starch catabolism are transported into the cytosol via chloroplast membrane transporters to enable metabolism in the cytosol. By analysing the extended model in the absence of light reactions such carbon interactions between chloroplast and cytosol can be studied further. Note that, in the dark, mitochondria are the main sources of ATP for cellular processes. The modified model can be used to investigate various routes involved in the exchange of ATP from mitochondria. Furthermore, the techniques described in this thesis can be used to identify various redox interactions between chloroplast, cytosol and mitochondria at night. A similar study describing the activity of the Calvin cycle in the absence of light was described in [63].

Experimental validation would immensely increase the value of the results presented in the thesis. This should include, in particular, the verification of two important predictions. Firstly, the role of G6P and PEP transporters in the transfer of ATP and reducing equivalents across the chloroplast membrane could be tested using experimental techniques. One possible starting point for such a study would be to follow the procedure described by Heineke *et al.* (1991) [148]. They used a subcellular fractionation technique to assay the concentration of reducing equivalents in the stroma and the cytosol for describing the activity of triose-phosphate/PGA and MAL/OAA shuttle mechanisms. A second, easier, approach would be to employ kinetic modelling techniques. Using the vast amount of kinetic information that is available in

public databases and literature sources, detailed kinetic models can be constructed to investigate the role of GPT and PPT. An example for a kinetic model constructed to investigate the role of triose-phosphate/PGA and MAL/OAA shuttle mechanisms in the exchange of ATP and reducing equivalents is available in [173].

Secondly, predictions of the localisation of enzymes made by analysing microarray data with RCCs can be tested using molecular biology techniques. One widely-used method involves tagging a gene coding for the enzyme of interest with fluorescent proteins such as GFP and expressing it in live plant cells. Localisation of the enzyme can then be determined using the confocal laser scanning microscopy technique. A recent study that illustrates the use of GFP fusions to determine the localisation of uncharacterised proteins is described in [193]. I have already initiated such an investigation, involving 24 carefully selected genes coding for reactions in the integrated model, in collaboration with the Plant Cell Biology Group at Oxford Brookes University, but experiments are not yet complete.

References

[1] H. Kitano, *Foundations of systems biology*, ch. Systems biology: toward system-level understanding of biological systems, pp. 1–29. The MIT Press, London, 2001.

[2] M. G. Poolman, H. E. Assmus, and D. A. Fell, "Applications of metabolic modelling to plant metabolism," *J. Exp. Bot.*, vol. 55, no. 400, pp. 1177–1186, 2004.

[3] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "KEGG for linking genomes to life and the environment," *Nucl. Acids Res.*, vol. 36, pp. D480–D484, 2008.

[4] R. Caspi, H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S. Y. Rhee, C. Tissier, P. Zhang, and P. D. Karp, "MetaCyc: a multiorganism database of metabolic pathways and enzymes," *Nucl. Acids Res.*, vol. 34, no. 1, pp. D511–516, 2006.

[5] A. Chang, M. Scheer, A. Grote, I. Schomburg, and D. Schomburg, "BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009.," *Nucl. Acids Res.*, vol. 37, pp. D588–D592, 2009. Database issue.

[6] A. Cornish-Bowden, *Fundamentals of enzyme kinetics*. Portland Press, London, 3rd ed., 2004.

[7] J. E. Bailey, "Mathematical modeling and analysis in biochemical engineering: Past accomplishments and future opportunities," *Biotechnol. Prog.*, vol. 14, pp. 8–20, 1998.

[8] B. O. Palsson, "The challenges of in silico biology," *Nat. Biotechnol.*, vol. 18, pp. 1147–1150, 2000.

[9] J. S. Edwards and B. O. Palsson, "Systems properties of the *Haemophilus influenzae* rd metabolic genotype," *J. Biol. Chem.*, vol. 274, no. 25, pp. 17410–17416, 1999.

[10] J. S. Edwards and B. O. Palsson, "The *E.coli* MG1655 in silico metabolic genotype: its definition, characteristics and capabilities," *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 5528–5533, 2000.

[11] C. H. Schilling, M. W. Covert, I. Famili, G. M. Church, J. S. Edwards, and B. O. Palsson, "Genome-scale metabolic model of *Helicobacter pylori* 26695," *J. Bacteriology*, vol. 184, no. 16, pp. 4582–4593, 2002.

[12] J. Forster, I. Famili, P. Fu, B. O. Palsson, and J. Nielsen, "Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network," *Genome Res.*, vol. 13, no. 2, pp. 244–253, 2003.

[13] M. G. Poolman, L. Miguet, L. J. Sweetlove, and D. A. Fell, "A genome-scale metabolic model of *Arabidopsis thaliana* and some of its properties," *Plant Physiol.*, vol. 151, no. 3, pp. 1570–1581, 2009.

[14] S. Schuster, T. Dandekar, and D. A. Fell, "Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering.," *Trends. Biotech.*, vol. 17, no. 2, pp. 53–60, 1999.

[15] D. A. Fell, *Biological networks: complex systems and interdisciplinary science*, vol. 3, ch. Metabolic networks, pp. 163–197. World Scientific, 2007.

[16] R. Heinrich and S. Schuster, *The regulation of cellular systems*. Chapman & Hall, London, 1996.

[17] C. H. Schilling, S. Schuster, B. O. Palsson, and R. Heinrich, "Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era," *Biotechnol. Prog.*, vol. 15, pp. 296–303, 1999.

[18] M. A. Savageau, *Biochemical systems analysis: a study of function and design in molecular biology*. Addison–Wesley, London, 1976.

[19] J.-H. S. Hofmeyr, "Steady-state modelling of metabolic pathways: a guide for the prospective simulator," *Comput. Appl. Biosci.*, vol. 2, no. 1, pp. 5–11, 1986.

[20] A. Cornish-Bowden, J.-H. S. Hofmeyr, and M. L. Cárdenas, "Stoicheiometric analysis in studies of metabolism," *Biochem. Soc. Trans.*, vol. 30, no. 2, pp. 43–46, 2002.

[21] D. C. Lay, *Linear algebra and its applications*. Addison–Wesley, Reading, MA, 3 ed., 2005.

[22] I. Famili and B. O. Palsson, "Systemic metabolic reactions are obtained by singular value decomposition of genome-scale stoichiometric matrices," *J. Theor. Biol.*, vol. 224, no. 1, pp. 87–96, 2003.

[23] B. O. Palsson, *Systems biology: properties of reconstructed networks*. Cambridge University Press, New York, 2006.

[24] C. H. Schilling and B. O. Palsson, "The underlying pathway structure of biochemical reaction networks," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 4193–4198, 1998.

[25] S. Klamt and J. Stelling, *System modeling in cellular biology: from concepts to nuts and bolts*, ch. Stoichiometric and constraint-based modeling, pp. 73–96. The MIT Press, London, 2006.

[26] M. W. Covert and B. O. Palsson, "Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*," *J. Biol. Chem.*, vol. 277, no. 31, pp. 28058–28064, 2002.

[27] F. Llaneras and J. Pico, "Stoichiometric modelling of cell metabolism," *J. Biosci. Bioeng.*, vol. 105, no. 1, pp. 1–11, 2008.

[28] M. G. Poolman, B. K. Bonde, A. Gevorgyan, H. H. Patel, and D. A. Fell, "Challenges to be faced in the reconstruction of metabolic networks from public databases," *Syst. Biol.*, vol. 153, no. 5, pp. 379–384, 2006.

[29] H. M. Sauro and B. P. Ingalls, "Computational analysis in biochemical networks: computational issues for software writers," *Biophys. Chem.*, vol. 109, pp. 1–15, 2004.

[30] A.-L. Barabasi and Z. N. Oltvai, "Network biology: Understanding the cell's functional organization," *Nat. Rev. Genet.*, vol. 5, pp. 101–113, 2004.

[31] R. Albert, "Scale-free networks in cell biology," *J. Cell Sci.*, vol. 118, no. 21, pp. 4947–4957, 2005.

[32] E. Almaas, A. Vazquez, and A.-L. Barabasi, *Biological networks: complex systems and interdisciplinary science*, vol. 3, ch. Scale-free networks in biology, pp. 163–197. World Scientific, 2007.

[33] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[34] A. Wagner and D. A. Fell, "The small world inside large metabolic networks," *Proc. Biol. Sci.*, vol. 268, no. 1478, pp. 1803–1810, 2001.

[35] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, p. 47, 2002.

[36] N. Lemke, F. Heredia, C. K. Barcellos, A. N. dos Reis, and J. C. M. Mombach, "Essentiality and damage in metabolic networks," *Bioinformatics*, vol. 20, no. 1, pp. 115–119, 2004.

[37] T. Pfeiffer, I. Sanchez-Valdenebro, J. C. Nuno, F. Montero, and S. Schuster, "METATOOL: for studying metabolic networks.," *Bioinformatics*, vol. 15, no. 3, pp. 251–257, 1999.

[38] S. Schuster, S. Klamt, W. Weckwerth, M. Moldenhauer, and T. Pfeiffer, "Use of network analysis of metabolic systems in bioengineering.," *Bioprocess Biosys. Eng.*, vol. 24, pp. 363–373, 2002.

[39] J. L. Reed and B. O. Palsson, "Genome-scale *in silico* models of *E.coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states.," *Genome Res.*, vol. 14, pp. 1797–1805, 2004.

[40] J. A. Papin, N. D. Price, and B. O. Palsson, "Extreme pathway lengths and reaction participation in genome-scale metabolic networks," *Genome Res.*, vol. 12, no. 12, pp. 1889–1900, 2002.

[41] B. K. Bonde, *Metabolism and bioinformatics: the relationship between metabolism and genome structure*. PhD thesis, Oxford Brookes University, 2006.

[42] H. Patel, *The structural analysis of metabolism on a genome scale*. PhD thesis, Oxford Brookes University, 2009.

[43] M. G. Poolman, C. Sebu, M. K. Pidcock, and D. A. Fell, "Modular decomposition of metabolic systems via null-space analysis," *J. Theor. Biol.*, vol. 249, no. 4, pp. 691–705, 2007.

[44] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, no. 25, pp. 14863 –14868, 1998.

[45] H. C. Causton, J. Quackenbush, and A. Brazma, *A biginner's guide to microarray gene expression data analysis*, ch. Clustering, pp. 98–112. Blackwell Publishing, Oxford, 2003 ed., 2003.

[46] B. J. T. Morgan and A. P. G. Ray, "Non-uniqueness and inversions in cluster analysis," *Applied Statistics*, vol. 44, no. 1, pp. 117–34, 1995.

[47] G. Perriere and M. Gouy, "WWW-Query: An on-line retrieval system for biological sequence banks," *Biochimie*, vol. 78, pp. 364–369, 1996.

[48] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software Version 4.0," *Mol. Biol. Evol.*, vol. 24, no. 8, pp. 1596–1599, 2007.

[49] S. Schuster, D. A. Fell, and T. Dandekar, "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks.," *Nature Biotech.*, vol. 18, pp. 326–332, 2000.

[50] S. Schuster, C. Hilgetag, J. H. Woods, and D. A. Fell, "Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucelotide metabolism.," *J. Math. Biol.*, vol. 45, pp. 153–181, 2002.

[51] J. A. Papin, N. D. Price, S. J. Wiback, D. A. Fell, and B. O. Palsson, "Metabolic pathways in the post-genomic era.," *Trends Biochem. Sci.*, vol. 28, no. 5, pp. 250–258, 2003.

[52] S. Schuster and C. Hilgetag, "On elementary flux modes in biochemical systems at steady state.," *J. Biol. Syst.*, vol. 2, no. 2, pp. 165–182, 1994.

[53] J. A. Papin, J. Stelling, N. D. Price, S. Klamt, S. Schuster, and B. O. Palsson, "Comparison of network-based pathway analysis methods," *Trends Biotechnol.*, vol. 22, no. 8, pp. 400–405, 2004.

[54] D. A. Fell, *Understanding the Control of Metabolism*. London: Portland Press, 1997.

[55] C. Wagner, "Nullspace approach to determine the elementary modes of chemical reaction systems," *J. Phys. Chem. B*, vol. 108, no. 7, pp. 2425–2431, 2004.

[56] J. Gagneur and S. Klamt, "Computation of elementary modes: a unifying framework and the new binary approach," *BMC Bioinformatics*, vol. 5, no. 175, 2004.

[57] S. Klamt and J. Stelling, "Combinatorial complexity of pathway analysis in metabolic networks," *Mol. Biol. Rep.*, vol. 29, no. 1-2, pp. 233–236, 2002.

[58] S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, and T. Dandekar, "Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *M. pneumoniae*," *Bioinformatics*, vol. 18, no. 2, pp. 351–361, 2002.

[59] J. Stelling, S. Klamt, B. Bettenbrock, S. Schuster, and E. D. Gilles, "Metabolic network structure determines key aspects of functionality and regulation," *Nature*, vol. 420, pp. 190–193, 2002.

[60] T. Cakir, C. S. Tacer, and K. O. Ulgen, "Metabolic pathway analysis of enzyme-deficient human red blood cells," *Biosystems*, vol. 78, no. 1-3, pp. 49–67, 2004.

[61] C. H. Schilling and B. O. Palsson, "Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis.," *J. Theor. Biol.*, vol. 203, no. 3, pp. 249–283, 2000.

[62] R. Carlson, D. A. Fell, and F. Srienc, "Metabolic pathway analysis of a recombinant yeast for rational strain development.," *Biotechnol. Bioeng.*, vol. 79, no. 2, pp. 121–134, 2002.

[63] M. G. Poolman, D. A. Fell, and C. A. Raines, "Elementary modes analysis of photosynthate metabolism in the chloroplast stroma," *Eur. J. Biochem*, vol. 270, pp. 430–439, 2003.

[64] R. Steuer, A. N. Nesi, A. R. Fernie, T. Gross, B. Blasius, and J. Selbig, "From structure to dynamics of metabolic pathways: application to the plant mitochondrial TCA cycle," *Bioinformatics*, vol. 23, pp. 1378–1385, 2007.

[65] J. C. Liao, S. Y. Hou, and Y. P. Chao, "Pathway analysis, engineering, and physiological considerations for redirecting central metabolism," *Biotechnol. Bioeng.*, vol. 52, pp. 129–140, 1996.

[66] C. H. Schilling, D. Letscher, and B. O. Palsson, "Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway oriented perspective.," *J. Theor. Biol.*, vol. 203, pp. 229–248, 2000.

[67] S. Klamt and J. Stelling, "Two approaches for metabolic pathway analysis?," *Trends Biochem. Sci.*, vol. 21, no. 2, pp. 64–69, 2003.

[68] J. A. Papin, N. D. Price, J. S. Edwards, and B. O. Palsson, "The genome-scale metabolic extreme pathway structure in *Haemophilus influenzae* shows significant network redundancy," *J. Theor. Biol.*, vol. 215, pp. 67–82, 2002.

[69] N. D. Price, J. A. Papin, and B. O. Palsson, "Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis," *Genome Res.*, vol. 12, no. 5, pp. 760–769, 2002.

[70] S. J. Wiback and B. O. Palsson, "Extreme pathway analysis of human red blood cell metabolism," *Biophys. J.*, vol. 83, no. 2, pp. 808–818, 2002.

[71] G. N. Stephanopoulus, A. A. Aristidou, and J. Nielsen, *Metabolic engineeering principles and methodologies*. Academic Press, London, 1998.

[72] S. Klamt, S. Schuster, and E. D. Gilles, "Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulphur metabolism," *Biotechnol. Bioeng.*, vol. 77, no. 7, pp. 734–750, 2002.

[73] W. Wiechert, M. Mollney, S. Petersen, and A. A. de Graaf, "A universal framework for 13C metabolic flux analysis," *Metab. Eng.*, vol. 3, no. 3, pp. 265–283, 2001.

[74] B. D. Follstad, R. R. Balcarcel, G. Stephanopoulos, and D. I. Wang, "Metabolic flux analysis of hybridoma continuous culture steady state multiplicity," *Biotechnol. Bioeng.*, vol. 63, no. 6, pp. 675–83, 1999.

[75] J. Schwender, J. Ohlrogge, and Y. Shachar-Hill, "Understanding flux in plant metabolic networks," *Curr. Opin. Plant Biol.*, vol. 7, pp. 309–317, 2004.

[76] C. Herwig and U. von Stockar, "A small metabolic flux model to identify transient metabolic regulations in *Saccharomyces cerevisiae*," *Bioprocess Biosyst. Eng.*, vol. 24, no. 6, pp. 395–403, 2002.

[77] A. Varma and B. O. Palsson, "Metabolic flux balancing: basic concepts, scientific and practical use," *Nat. Biotechnol.*, vol. 12, pp. 994–998, 1994.

[78] J. S. Edwards, M. W. Covert, and B. O. Palsson, "Metabolic modelling of microbes: the flux-balance approach," *Environ. Microbiol.*, vol. 4, no. 3, pp. 133–140, 2002.

[79] K. J. Kauffman, P. Prakash, and J. S. Edwards, "Advances in flux balance analysis," *Curr. Opin. Biotechnol.*, vol. 14, no. 5, pp. 491–496, 2003.

[80] J. S. Edwards, R. U. Ibarra, and B. O. Palsson, "In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data," *Nat. Biotechnol.*, vol. 19, no. 2, pp. 125–130, 2001.

[81] D. Segre, D. Vitkup, and G. M. Church, "Analysis of optimality in natural and perturbed metabolic networks," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 23, pp. 15112–15117, 2002.

[82] J.-J. Zhong, "Plant cell culture for production of paclitaxel and other taxanes," *J. Biosci. Bioeng.*, vol. 94, no. 6, pp. 591–599, 2002.

[83] H. C. Causton, J. Quackenbush, and A. Brazma, *A biginner's guide to microarray gene expression data analysis*, ch. Image processing, normalisation and data transformation, pp. 40–70. Blackwell Publishing, Oxford, 2003 ed., 2003.

[84] M. M. Babu, *Computational Genomics: Theory and Application*, ch. An introduction to microarray data analysis, pp. 225–249. Horizon Scientific Press, Norwich, UK, 2004.

[85] D. J. Craigon, N. James, J. Okyere, J. Higgins, J. Jotham, and S. May, "NASCArrays: A repository for microarray data generated by NASC's transcriptomics service," *Nucleic Acids Res.*, vol. 32, pp. Database issue D575–D577, 2004.

[86] P. Zhang, H. Foerster, C. P. Tissier, L. Mueller, S. Paley, P. D. Karp, and S. Y. Rhee, "MetaCyc and AraCyc. Metabolic pathway databases for plant research," *Plant Physiol.*, vol. 138, pp. 27–37, 2005.

[87] A. Pettinen, T. Aho, O.-P. Smolander, T. Manninen, A. Saarinen, K.-L. Taattola, O. Yli-Harja, and M.-L. Linne, "Simulation tools for biochemical networks: evaluation of performance and usability," *Bioinformatics*, vol. 21, no. 3, pp. 357–363, 2005.

[88] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer, "COPASI - a COmplex PAthway SImulator," *Bioinformatics*, vol. 22, pp. 3067–3074, 2006.

[89] P. Mendes, "Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3," *Trends Biochem. Sci.*, vol. 22, pp. 361–363, 1997.

[90] R. Schwarz, P. Musch, A. von Kamp, B. Engels, H. Schirmer, S. Schuster, and T. Dandekar, "YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities," *BMC Bioinformatics*, vol. 6, no. 135, pp. 1–12, 2005.

[91] S. Klamt, J. Stelling, M. Ginkel, and E. D. Gilles, "FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps," *Bioinformatics*, vol. 19, no. 2, pp. 261–269, 2003.

[92] H. M. Sauro, M. Hucka, A. Finney, C. Wellock, H. Bolouri, J. Doyle, and H. Kitano, "Next generation simulation tools: the Systems Biology Workbench and BioSPICE Integration.," *OMICS*, vol. 7, no. 4, pp. 355–372, 2003.

[93] H. M. Sauro, "SCAMP: a general-purpose metabolic simulator and metabolic control analysis program," *Comp. Appl. Biosci.*, vol. 9, no. 4, pp. 441–450, 1993.

[94] B. G. Olivier, J. M. Rohwer, and J.-H. S. Hofmeyr, "Modelling cellular systems with PySCeS.," *Bioinformatics*, vol. 21, no. 4, pp. 560–561, 2005.

[95] M. G. Poolman, "ScrumPy - metabolic modelling with Python," *IET Syst Biol.*, vol. 153, no. 5, pp. 375–378, 2006.

[96] M. G. Poolman, K. V. Venkatesh, M. K. Pidcock, and D. A. Fell, "A method for the determination of flux in elementary modes, and its application to *Lactobacillus rhamnosus*," *Biotechnol. Bioeng.*, vol. 88, no. 5, pp. 601–612, 2004.

[97] M. Hucka, A. Finney, H. M. Sauro, and *et al.*, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical networks," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[98] S. Klamt, J. Saez-Rodriguez, and E. D. Gilles, "Structural and functional analysis of cellular networks with CellNetAnalyzer," *BMC Syst. Biol.*, vol. 1, no. 2, pp. 1–13, 2007.

[99] M. Lutz and D. Ascher, *Learning Python*. O'Reilly & Associates, CA, 1st ed., 1999.

[100] M. Lutz, *Programming Python*. O'Reilly & Associates, CA, 2nd ed., 2001.

[101] L. Prechelt, "An empirical comparison of seven programming languages," *Computer*, vol. 30, no. 10, pp. 23–29, 2000.

[102] B. G. Olivier, J. M. Rohwer, and J.-H. S. Hofmeyr, "Modelling cellular processes with Python and SciPy," *Mol. Biol. Rep.*, vol. 29, pp. 249–254, 2002.

[103] L. J. Sweetlove and A. R. Fernie, "Regulation of metabolic networks: understanding metabolic complexity in the systems biology era," *New Phytol.*, vol. 168, pp. 9–24, 2005.

[104] L. J. Sweetlove, D. A. Fell, and A. R. Fernie, "Getting to grips with the plant metabolic network," *Biochem. J.*, vol. 409, no. 1, pp. 27–41, 2008.

[105] F. B. Salisbury and C. W. Ross, *Plant physiology*, ch. Plant physiology and plant cells, pp. 3–26. Wadsworth Publishing Company, CA, 4th ed., 1992.

[106] H. Mohr and P. Schopfer, *Plant physiology*, ch. The cell as a morphological system, pp. 22–37. Springer-Verlag, Berlin, 1995.

[107] H. Beevers, *Compartmentation of plant metabolism in non–photosynthetic tissues*, ch. Metabolic compartmentation in plant cells, pp. 1–23. Cambridge University Press, Cambridge, 1991.

[108] M. J. Emes and D. T. Dennis, *Plant metabolism*, ch. Regulation by compartmentation, pp. 69–80. Addison–Wesley Longman, Harlow, 2nd ed., 1997.

[109] D. L. Nelson and M. M. Cox, *Principles of biochemistry*, ch. Oxidative phosphorylation and photophosphorylation, pp. 707–772. W. H. Freeman and company, NY, 5 ed., 2008.

[110] H. W. Heldt and F. Sauer, "The inner membrane of the chloroplast envelope as the site of specific metabolite transport," *Biochim. Biophys. Acta*, vol. 234, no. 1, pp. 83–91, 1971.

[111] D. A. Walker, "Excited leaves," *New Phytologist*, vol. 121, no. 3, pp. 325–345, 1992.

[112] J. F. Allen, "Photosynthesis of ATP - electrons, proton pumps, rotors, and poise," *Cell*, vol. 110, no. 3, pp. 273–276, 2002.

[113] G. N. Johnson, "Cyclic electron transport in C_3 plants: fact or artefact?," *J. Exp. Bot.*, vol. 56, no. 411, pp. 407–416, 2005.

[114] P.-A. Albertsson, "A quantitative model of the domain structure of the photosynthetic membrane," *Trends Plant Sci.*, vol. 6, no. 8, pp. 349–354, 2001.

[115] D. M. Kramer, J. A. Cruz, and A. Kanazawa, "Balancing the central roles of the thylakoid proton gradient," *Trends Plant Sci.*, vol. 8, no. 1, pp. 27–32, 2003.

[116] F. B. Salisbury and C. W. Ross, *Plant physiology*, ch. Carbon dioxide fixation and carbohydrate synthesis, pp. 225–248. Wadsworth Publishing Company, CA, 4th ed., 1992.

[117] D. L. Nelson and M. M. Cox, *Principles of biochemistry*, ch. Carbohydrate biosynthesis in plants and bacteria, pp. 773–804. W. H. Freeman and company, NY, 5 ed., 2008.

[118] M. J. Emes and H. E. Neuhaus, "Metabolism and transport in non–photosynthetic plastids," *J. Exp. Bot.*, vol. 48, no. 317, pp. 1995–2005, 1997.

[119] H. E. Neuhaus and M. J. Emes, "Nonphotosynthetic metabolism in plastids," *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, vol. 51, no. 1, pp. 111–140, 2000.

[120] L. E. Anderson, "Light/dark modulation of enzyme activity in plants," *Adv. Bot. Res.*, vol. 12, pp. 1–46, 1986.

[121] N. Zhang and A. R. Portis, "Mechanism of light regulation of Rubisco: A specific role for the larger Rubisco activase isoform involving reductive activation by thioredoxin-f," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 16, pp. 9438–9443, 1999.

[122] C. Schnarrenberger, A. Flechner, and W. Martin, "Enzymatic evidence for a complete oxidative pentose phosphate pathway in chloroplasts and an incomplete pathway in the cytosol of spinach leaves," *Plant Physiol.*, vol. 108, no. 2, pp. 609–614, 1995.

[123] F. B. Salisbury and C. W. Ross, *Plant Physiology*, ch. Respiration, pp. 266–288. Wadsworth Publishing Company, CA, 4th ed., 1992.

[124] N. J. Kruger and A. von Schaewen, "The oxidative pentose phosphate pathway: structure and organisation," *Curr. Opin. Plant Biol.*, vol. 6, no. 3, pp. 236–246, 2003.

[125] S. C. Zeeman, S. M. Smith, and A. M. Smith, "The breakdown of starch in leaves," *New Phytol.*, vol. 163, no. 2, pp. 247–261, 2004.

[126] D. L. Nelson and M. M. Cox, *Principles of biochemistry*, ch. Glycolysis, gluconeogenesis, and the pentose phosphate pathway, pp. 527–568. W. H. Freeman and company, NY, 5 ed., 2008.

[127] W. C. Plaxton, "The organization and regulation of plant glycolysis," *Ann. Rev. Plant Physiol. Plant Mol. Biol.*, vol. 47, pp. 185–214, 1996.

[128] C. V. Givan, "Evolving concepts in plant glycolysis : two centuries of progress," *Biol. Rev.*, vol. 74, no. 3, pp. 277–309, 1999.

[129] L. D. Gottlieb, "Conservation and duplication of isozymes in plants," *Science*, vol. 216, no. 4544, pp. 373–380, 1982.

[130] N. F. Weeden and J. F. Wendel, *Isozymes in plant biology*, ch. Genetics of plant isozymes, pp. 46–73. Chapman & Hall, London, 1990.

[131] S.-J. S. Sung, D.-P. Xu, C. M. Galloway, and C. C. Black, "A reassessment of glycolysis and gluconeogenesis in higher plants," *Physiol. Plant.*, vol. 72, no. 3, pp. 650–654, 1988.

[132] K. Fischer and A. Weber, "Transport of carbon in non-green plastids.," *Trends Plant Sci.*, vol. 7, no. 8, pp. 345–351, 2002.

[133] A. R. Fernie, F. Carrari, and L. J. Sweetlove, "Respiratory metabolism: glycolysis, the TCA cycle and mitochondrial electron transport," *Curr. Opin. Plant Biol.*, vol. 7, pp. 254–261, 2004.

[134] M. Saraste, "Oxidative phosphorylation at the fin de siecle," *Science*, vol. 283, no. 5407, p. 1488, 1999.

[135] C. G. Bowsher and A. K. Tobin, "Compartmentation of metabolism within mitochondria and plastids," *J. Exp. Bot.*, vol. 52, no. 356, pp. 513–527, 2001.

[136] M. W. Gray, G. Burger, and B. F. Lang, "Mitochondrial evolution," *Science*, vol. 283, no. 5407, pp. 1476–1481, 1999.

[137] G. Duby and M. Boutry, "Mitochondrial protein import machinery and targeting information," *Plant Science*, vol. 162, no. 4, pp. 477 –490, 2002.

[138] D. C. Logan, "The mitochondrial compartment," *J. Exp. Bot.*, vol. 57, no. 6, pp. 1225–1243, 2006.

[139] H. W. Heldt and U.-I. Flügge, *The biochemistry of plants*, vol. 12, ch. Subcellular transport of metabolites in plant cells, pp. 49–85. Academic Press, New York, 1987.

[140] U.-I. Flügge and H. W. Heldt, "Metabolite translocators of the choroplast envelope," *Annu. Rev. Plant Physiol. Mol. Biol.*, vol. 42, pp. 129–144, 1991.

[141] U.-I. Flügge, "Phosphate translocators in plastids," *Ann. Rev. Plant Physiol. Plant Mol. Biol.*, vol. 50, pp. 27–45, 1999.

[142] A. P. M. Weber, J. Schneidereit, and L. M. Voll, "Using mutants to probe the *in vivo* function of plastid envelope membrane metabolite transporters," *J. Exp. Bot.*, vol. 55, no. 400, pp. 1231–1244, 2003.

[143] A. P. M. Weber, R. Schwacke, and U.-I. Flügge, "Solute transporters of the plastid envelope membrane," *Annu. Rev. Plant Biol.*, vol. 56, pp. 133–164, 2005.

[144] H. E. Neuhaus and R. Wagner, "Solute pores, ion channels, and metabolite transporters in the outer and inner envelope membranes of higher plant plastids.," *Biochim.Biophys.Acta*, vol. 1456, pp. 207–323, 2000.

[145] K. Fischer, B. Kammerer, M. Gutensohn, B. Arbinger, A. P. M. Weber, R. E. Hausler, and U.-I. Flügge, "A new class of plastidic phosphate translocators: a putative link between primary and secondary metabolism by the phosphoenolpyruvate/phosphate antiporter," *Plant Cell*, vol. 9, pp. 453–462, 1997.

[146] A. P. M. Weber, "Solute transporters as connecting elements between cytosol and plastid stroma," *Curr. Opin. Plant Biol.*, vol. 7, pp. 247–253, 2004.

[147] U. Heber, "Energy transfer within leaf cells," in *Proc. Int. Congr. Photosynth.* (M. Avron, ed.), vol. 2, pp. 1335–1348, Elsevier Scientific Publishing, Amsterdam, 1975.

[148] D. Heineke, B. Riens, H. Große, P. Hofereichter, U. Peter, U.-I. Flügge, and H. W. Heldt, "Redox transfer across the inner chloroplast envelope membrane.," *Plant Physiol.*, vol. 95, pp. 1131–1137, 1991.

[149] M. H. N. Hoefnagel, O. K. Atkin, and J. T. Wiskich, "Interdependence between chloroplasts and mitochondria in the light and the dark," *Biochim. Biophys. Acta*, vol. 1366, no. 3, pp. 235 – 255, 1998.

[150] R. Scheibe, "$NADP^+$–malate dehydrogenase in C_3–plants: regulation and role of a light–activated enzyme," *Physiol. Plant.*, vol. 71, no. 3, pp. 393–400, 1987.

[151] S. Kromer, "Respiration during photosynthesis," *Annu. Rev. Plant. Physiol. Plant Mol. Biol.*, vol. 46, pp. 45–70, 1995.

[152] O. K. Atkin, A. H. Millar, P. Gardestrom, and D. A. Day, *Photosynthesis: physiology and metabolism*, vol. 9, ch. Photosynthesis, carbohydrate metabolism and respiration in leaves of higher plants, pp. 154–170. Kluwer Academic Publishers, Dordrecht, 2000.

[153] H. W. Heldt, "Adenine nucleotide translocation in spinach chloroplasts," *FEBS Lett.*, vol. 5, no. 1, pp. 11–14, 1969.

[154] U. Heber and K. A. Santarius, "Direct and indirect transfer of ATP and ADP across the chloroplast envelope," *Z Naturforsch B.*, vol. 25, no. 7, pp. 718–728, 1970.

[155] H. W. Heldt and U.-I. Flügge, *Plant organelles: compartmentation of metabolism in photosynthetic cells*, ch. Metabolite transport in plant cells, pp. 21–46. Cambridge University Press, Cambridge, 1992.

[156] M. Stitt, *Plant metabolism*, ch. The flux of carbon between the chloroplast and cytoplasm, pp. 382–400. Addison–Wesley Longman, Harlow, 1997.

[157] H. E. Neuhaus, G. Henrichs, and R. Scheibe, "Characterization of glucose-6-phosphate incorporation into starch by isolated intact cauliflower-bud plastids," *Plant Physiol.*, vol. 101, no. 2, pp. 573–578, 1993.

[158] H. E. Neuhaus and N. Schulte, "Starch degradation in chloroplasts isolated from C_3 or CAM (crassulacean acid metabolism) induced *Mesembrianthenum crytallinum* L.," *Biochem. J.*, vol. 318, pp. 945–953, 1996.

[159] M. Stitt, R. M. Lilley, and H. W. Heldt, "Adenine nucleotide levels in the cytosol, chloroplasts, and mitochondria of wheat leaf protoplasts," *Plant Physiol.*, vol. 70, no. 4, pp. 971–977, 1982.

[160] A. Pradet and P. Raymond, "Adenine nucleotide ratios and adenylate energy charge in energy metabolism," *Ann. Rev. Plant Physiol.*, vol. 34, no. 1, pp. 199–224, 1983.

[161] I. Hanning and H. W. Heldt, "On the function of mitochondrial metabolism during photosynthesis in spinach (*Spinacia oleracea* L.) leaves (partitioning between respiration and export of redox equivalents and precursors for nitrate assimilation products).," *Plant Physiol.*, vol. 103, no. 4, pp. 1147–1154, 1993.

[162] K. Padmasree, L. Padmavathi, and A. S. Raghavendra, "Essentiality of mitochondrial oxidative metabolism for photosynthesis: optimization of carbon assimilation and protection against photoinhibition," *Crit. Rev. Biochem. Mol. Biol.*, vol. 37, no. 2, pp. 71–119, 2002.

[163] A. S. Raghavendra and K. Padmasree, "Beneficial interactions of mitochondrial metabolism with photosynthetic carbon assimilation," *Trends Plant Sci.*, vol. 8, no. 11, pp. 546–553, 2003.

[164] K. K. Niyogi, "Safety valves for photosynthesis," *Curr. Opin. Plant Biol.*, vol. 3, no. 6, pp. 455–460, 2000.

[165] M.-L. Oelze, A. Kandlbinder, and K.-J. Dietz, "Redox regulation and overreduction control in the photosynthesizing cell: Complexity in redox regulatory networks," *Biochim. Biophys. Acta*, vol. 1780, no. 11, pp. 1261–1272, 2008.

[166] B. Chance, D. Garfinkel, J. Higgins, and B. J. Hess, "A solution for the equations representing interaction between glycolysis and respiration in ascites tumor cells," *J. Biol. Chem.*, vol. 235, no. 8, pp. 2426–2439, 1960.

[167] D. Garfinkel and B. Hess, "Metabolic control mechanisms VII. a detailed computer model of the glycolytic pathway in ascites cells.," *J. Biol. Chem*, vol. 239, no. 4, pp. 971–983, 1964.

[168] H. Kacser and J. Burns, "The control of flux," *Symp. Soc. Exp. Biol.*, vol. 27, pp. 65–104, 1973.

[169] A. Laisk, "Mathematical model of photosynthesis and photorespiration. reversible phosphoribulokinase reaction," *Biofizika*, vol. 18, no. 4, pp. 637–642, 1973.

[170] J. H. M. Thornley, "Light fluctuations and photosynthesis," *Ann. Bot.*, vol. 38, pp. 363–373, 1974.

[171] J. Milstein and H. J. Bremermann, "Parameter identification of the Calvin photosynthesis cycle," *J. Math. Biol.*, vol. 7, pp. 99–116, 1979.

[172] V. Kaitala, P. Hari, E. Vapaavuori, and R. Salminen, "A dynamic model for photosynthesis," *Ann. Bot.*, vol. 50, pp. 385–396, 1982.

[173] C. Giersch, U. Heber, and G. H. Krause, *Plant Membrane Transport: Current Conceptual Issues*, ch. ATP transfer from chloroplasts to the cytosol of the leaf cells during photosynthesis and its effect on leaf metabolism, pp. 65–80. Elsevier, North-Holland Biomedical Press, 1980.

[174] I. E. Woodrow, "Control of the rate of photosynthetic carbon dioxide fixation," *Biochim. Biophy. Acta*, vol. 851, pp. 181–192, 1986.

[175] G. Pettersson and U. Ryde-Pettersson, "A mathematical model of the Calvin photosynthesis cycle," *Eur. J. Biochem.*, vol. 175, pp. 661–672, 1988.

[176] A. Laisk, H. Eichelmann, A. Eatheall, and D. A. Walker, "A mathematical model of carbon metabolism in photosynthesis: Difficulties in explaining oscillations by Fructose 2,6-bisphosphate regulation," *Proc. R. Soc. Lond. B*, vol. 237, pp. 389–415, 1989.

[177] C. Giersch, M. N. Sivak, and D. A. Walker, "A mathematical skeleton model of photosynthetic oscillations," *Proc. R. Soc. Lond. B*, vol. 245, pp. 77–83, 1991.

[178] M. G. Poolman, *Computer modelling applied to the Calvin cycle*. PhD thesis, Oxford Brookes University, 1999.

[179] M. G. Poolman and D. A. Fell, "Modelling photosynthesis and its control," *J. Exp. Bot*, vol. 51, pp. 319–328, 2000.

[180] M. G. Poolman and D. A. Fell, "Modelling and experimental evidence for two seperate steady-states in the photosynthetic Calvin cycle.," in *BioThermoKinetics 2000 Animating the Cellular Map* (J.-H. S. Hofmeyr, J. M. Rohwer, and J. L. Snoep, eds.), pp. 249–254, Stellenbosch University Press, 2000.

[181] H. E. Assmus, *Modelling Carbohydrate Metabolism in Potato Tuber Cell*. PhD thesis, Oxford Brookes University, 2005.

[182] T. J. Avenson, A. Kanazawa, J. A. Cruz, K. Takizawa, W. E. Ettinger, and D. M. Kramer, "Integrating the proton circuit into photosynthesis: progress and challenges," *Plant, Cell and Environment*, vol. 28, pp. 97–109, 2005.

[183] C. Kaleta, L. F. de Figueiredo, and S. Schuster, "Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns," *Genome Res.*, vol. 19, no. 10, pp. 1872–1883, 2009.

[184] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[185] H. C. Causton, J. Quackenbush, and A. Brazma, *Microarray gene expression data analysis: a beginner's guide*. Blackwell Publishing, Oxford, 2003.

[186] J. Ihmels, R. Levy, and N. Barkai, "Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*," *Nature Biotech.*, vol. 22, no. 1, pp. 86–92, 2004.

[187] S. Persson, H. Wei, J. Milne, G. P. Page, and C. R. Somerville, "Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets," *Proc. Natl. Acad. Sci. U.S.A*, vol. 102, pp. 8633–8638, 2005.

[188] H. Wei, S. Persson, T. Mehta, V. Srinivasasainagendra, L. Chen, G. P. Page, C. R. Somerville, and A. Loraine, "Transcriptional coordination of the metabolic network in *Arabidopsis*," *Plant Physiol.*, vol. 142, no. 2, pp. 762–774, 2006.

[189] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus.," *Nat. Rev. Genet.*, vol. 7, no. 1, pp. 55–65, 2006.

[190] M. Menges, R. Doczi, L. Okresz, P. Morandini, L. Mizzi, M. Soloviev, J. A. H. Murray, and L. Bogre, "Comprehensive gene expression atlas for the *Arabidopsis* map kinase signalling pathways," *New Phytol.*, vol. 179, no. 3, pp. 643–662, 2008.

[191] B. Usadel, T. Obayashi, M. Mutwil, F. M. Giorgi, G. W. Bassel, M. Tanimoto, A. Chow, D. Steinhauser, S. Persson, and N. J. Provart, "Co-expression tools for plant biology: opportunities for hypothesis generation and caveats," *Plant Cell Environ.*, vol. 32, no. 12, pp. 1633–1651, 2009.

[192] T. P. J. Dunkley, P. Dupree, R. B. Watson, and K. S. Lilley, "The use of isotope-coded affinity tags (ICAT) to study organelle proteomes in *Arabidopsis thaliana*.," *Biochem. Soc. Trans.*, vol. 32, no. 3, pp. 520–523, 2004.

[193] T. P. J. Dunkley, S. Hester, I. P. Shadforth, J. Runions, T. Weimar, S. L. Hanton, J. L. Griffin, C. Bessant, F. Brandizzi, C. Hawes, R. B. Watson, P. Dupree, and

K. S. Lilley, “Mapping the *Arabidopsis* organelle proteome.,” *Proc. Natl. Acac. Sci. USA*, vol. 103, no. 17, pp. 6518–6523, 2006.

[194] M. R. Hanson and R. H. Kohler, “Gfp imaging: methodology and application to investigate cellular compartmentation in plants,” *J. Exp. Bot.*, vol. 52, no. 356, pp. 529–539, 2001.

[195] J. L. Heazlewood, R. E. Verboom, J. Tonti-Filippini, I. Small, and A. H. Millar, “SUBA: the Arabidopsis Subcellular Database,” *Nucleic Acids Res.*, vol. 00, pp. 1–6, 2006.

[196] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, “Locating proteins in the cell using TargetP, SignalP, and related tools,” *Nat. Protoc.*, vol. 2, pp. 953–971, 2007.

[197] I. Small, N. Peeters, F. Legeai, and C. Lurin, “Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences,” *Proteomics*, vol. 4, no. 6, pp. 1581–90, 2004.

[198] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano, “Extensive feature detection of N-terminal protein sorting signals,” *Bioinformatics*, vol. 18, no. 2, pp. 298–305, 2002.

[199] S. Hua and Z. Sun, “Support vector machine approach for protein subcellular localization prediction,” *Bioinformatics*, vol. 17, no. 8, pp. 721–728, 2001.

[200] M. G. Claros and P. Vincens, “Computational method to predict mitochondrially imported proteins and their targeting sequences,” *Eur. J. Biochem.*, vol. 241, no. 3, pp. 779–786, 1996.

[201] C. Guda, E. Fahy, and S. Subramaniam, “MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins,” *Bioinformatics*, vol. 20, no. 11, pp. 1785–1794, 2004.

[202] O. Emanuelsson, A. Elofsson, G. von Heijne, and S. Cristobal, “In silico prediction of the peroxisomal proteome in fungi, plants and animals,” *J. Mol. Biol.*, vol. 330, pp. 443–456, 2003.

[203] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, and K. Nakai, “WoLF PSORT: protein localization predictor,” *Nucleic Acids Res.*, vol. 35, pp. 1–3, 2007.

[204] J. L. Heazlewood, R. E. Verboom, J. Tonti-Filippini, and A. H. Millar, “Combining experimental and predicted datasets for determination of the

subcellular location of proteins in *Arabidopsis*," *Plant Physiol.*, vol. 139, no. 2, pp. 598–609, 2005.

[205] J. L. Reed, I. Famili, I. Thiele, and B. O. Palsson, "Towards multidimensional genome annotation," *Nat. Rev. Genet.*, vol. 7, pp. 130–141, 2006.

[206] S. Selvarasu, I. A. Karimi, G.-H. Ghim, and D.-Y. Lee, "Genome-scale modeling and *in silico* analysis of mouse cell metabolic network," *Mol. BioSyst.*, vol. 6, pp. 152–161, 2010.

Appendices

APPENDIX A

Model of photosynthetic light reactions in *.spy* format

```
#####################################################################
######## MODEL OF PHOTOSYNTHETIC ELECTRON TRANSPORT CHAIN #########
#####################################################################

Structural()
External(H2O_lum, Photon, O2, Proton_str)

PS2_lum:
4Photon + 2H2O_lum -> O2 + 4Proton_lum + 4e_hi ~

Q_lum:
e_hi + 2Proton_str +  2Q_o_lum ->  2Q_r_lum ~

PQ_lum:
Q_r_lum  + 2 PQ_lum ->  2PQH2_lum +  Q_o_lum ~

CytB6_lum:
2PQH2_lum +  CytB6_o_lum  -> 2PQ_lum +  CytB6_r_lum + 4Proton_lum ~

PC_lum:
CytB6_r_lum + PC_o_lum -> CytB6_o_lum + PC_r_lum ~

PS1_lum:
5Photon + PC_r_lum + FD_o_str -> PC_o_lum + FD_r_str ~

NADPRe_str:
FD_r_str + NADP_str +  Proton_str -> NADPH_str  + FD_o_str ~

Cyclic_lum:
FD_r_str  +  Q_r_lum -> FD_o_str  + Q_o_lum ~

ATPSy_str:
14Proton_lum + 3ADP_str + 3Pi_str ->  14Proton_str + 3ATP_str ~

ADPSy_str:
ATP_str -> ADP_str + Pi_str + x_ATPWork ~

NADPHOx_str:
NADPH_str -> NADP_str + Proton_str + x_NADPHWork ~

#####################################################################
```

Figure A.1 – Model of light reactions representing photosynthetic electron transport chain in *.spy* format. The model is described in Section 2.2.1.1 and illustrated in Figure 2.3. See List of Abbreviations for metabolite and reaction abbreviations. '_lum' and '_str'represent metabolites and reactions localised in lumen and stroma, respectively. '_o' and '_r' indicate the oxidised and reduced state of a metabolite, respectively.

APPENDIX B

Model of Calvin cycle in *.spy* format

```
#########################################################################
#####################  MODEL OF CALVIN CYCLE  ###########################
#########################################################################

Structural()

External(CO2, Starch_str, PGA_cyt, GAP_cyt, NADPH_str, NADP_str)
External(G6P_cyt, PEP_cyt, DHAP_cyt, Pi_cyt, MAL_cyt, OAA_cyt)

#########################################################################
########################   CO2 ASSIMILATION   ###########################
#########################################################################

Rubisco_str:
CO2 + RuBP_str -> 2 PGA_str ~

PGK_str:
PGA_str + ATP_str <> BPGA_str + ADP_str ~

GAPDH_str:
BPGA_str + NADPH_str <> NADP_str + GAP_str + Pi_str ~

TPI_str:
GAP_str <> DHAP_str ~

Ald1_str:
DHAP_str + GAP_str <>  FBP_str ~

FBPase_str:
FBP_str -> F6P_str + Pi_str ~

#########################################################################
########################   STARCH SYNTHESIS  ############################
#########################################################################

PGI_str:
F6P_str <> G6P_str ~

PGM_str:
G6P_str <> G1P_str ~

StSynth_str:
G1P_str + ATP_str -> ADP_str + 2 Pi_str + Starch_str ~

#########################################################################
#######################    STARCH DEGRADATION ###########################
#########################################################################

StPase_str:
Starch_str + Pi_str -> G1P_str ~
```

```
TKL1_str:
F6P_str + GAP_str <> E4P_str + X5P_str ~

Ald2_str:
E4P_str + DHAP_str <> SBP_str ~

SBPase_str:
SBP_str -> S7P_str + Pi_str ~

TKL2_str:
GAP_str + S7P_str <> X5P_str + R5P_str ~

R5Piso_str:
R5P_str <> Ru5P_str ~

X5Piso_str:
X5P_str <> Ru5P_str ~

Ru5Pk_str:
Ru5P_str + ATP_str -> RuBP_str + ADP_str ~

#####################################################################
#################   GLYCOLYTIC REACTIONS OF CHLOROPLAST #############
#####################################################################

PGlyM_str:
PGA_str <> PGA2_str ~

Eno_str:
PGA2_str <> PEP_str ~

PK_str:
PEP_str + ADP_str ->  PYR_str + ATP_str ~

ME_str:
PYR_str  +  CO2 + NADPH_str <> MAL_str + NADP_str ~

NADPMDH_str:
MAL_str + NADP_str <> OAA_str + NADPH_str ~

#####################################################################
################ TRANSPORT REACTIONS OF CHLOROPLAST #################
#####################################################################

TX_PGA_str:
PGA_str + Pi_cyt -> Pi_str + PGA_cyt ~

TX_GAP_str:
GAP_str + Pi_cyt -> Pi_str + GAP_cyt ~
```

```
TX_DHAP_str:
DHAP_str + Pi_cyt -> Pi_str + DHAP_cyt ~

TX_MAL_str:
MAL_str -> MAL_cyt ~

TX_OAA_str:
OAA_str -> OAA_cyt ~

TX_G6P_str:
G6P_str + Pi_cyt -> Pi_str + G6P_cyt ~

TX_PEP_str:
PEP_str + Pi_cyt -> Pi_str + PEP_cyt ~

#####################################################################
########################## SINK REACTIONS ###########################
#####################################################################

ATPase_str:
ADP_str + Pi_str -> ATP_str ~

#####################################################################
```

Figure B.1 – Model of the Calvin cycle described by Poolman *et al.* [63] extended to include 'glycolytic' reactions of the chloroplast. The Calvin cycle is reviewed in Section 2.2.1.2 and illustrated in Figure 2.4. See List of Abbreviations for metabolite and reaction abbreviations. '_cyt' and '_str' indicate metabolites localised in cytosol and stroma, respectively.

APPENDIX C

Model of the glycolytic reactions of cytosol

```
#####################################################################
############ MODEL OF GLYCOLYTIC REACTIONS OF CHLOROPLAST #########
#####################################################################


Structural()

External(Sucrose, CO2_cyt, ATP_cyt, ADP_cyt, PPi_cyt, Pi_cyt)
External(MAL_str,  PEP_str, G6P_str, DHAP_str, GAP_str, PGA_str)
External(Pi_str, OAA_str, PYR_str, NADH_cyt, NAD_cyt)


#####################################################################
###################### SUCROSE METABOLISM ###########################
#####################################################################

Suc_EX:
Sucrose_cyt -> Sucrose ~

PGI_cyt:
G6P_cyt <> F6P_cyt ~

PGM_cyt:
G6P_cyt <> G1P_cyt ~

SuSyn_cyt:
F6P_cyt + UDPG_cyt <> Sucrose_cyt + UDP_cyt + Pi_cyt ~

UGPase_cyt:
UDPG_cyt + PPi_cyt <> G1P_cyt + UTP_cyt ~

NDPK_cyt:
UTP_cyt  <> UDP_cyt ~

PFK_cyt:
F6P_cyt + ATP_cyt -> FBP_cyt + ADP_cyt ~

PFP_cyt:
F6P_cyt + PPi_cyt <> FBP_cyt + Pi_cyt ~

Ald1_cyt:
FBP_cyt <> GAP_cyt + DHAP_cyt ~

TPI_cyt:
DHAP_cyt <> GAP_cyt ~

GAPDHP_cyt:
GAP_cyt + NAD_cyt + Pi_cyt <> BPGA_cyt + NADH_cyt ~

PGK_cyt:
BPGA_cyt + ADP_cyt <> PGA_cyt + ATP_cyt ~
```

```
GAPDH_cyt:
GAP_cyt + NAD_cyt  -> PGA_cyt + NADH_cyt ~

PGlyM_cyt:
PGA_cyt <> PGA2_cyt ~

Eno_cyt:
PGA2_cyt <> PEP_cyt ~

PK_cyt:
PEP_cyt + ADP_cyt -> PYR_cyt + ATP_cyt ~

PEPC_cyt:
PEP_cyt + CO2_cyt -> OAA_cyt + Pi_cyt ~

NADMDH_cyt:
OAA_cyt + NADH_cyt -> MAL_cyt + NAD_cyt ~

#########################################################################
######################## TRANSPORT REACTIONS ############################
#########################################################################

TX_PYR_cyt:
PYR_cyt -> PYR_str ~

TX_PEP_cyt:
PEP_str + Pi_cyt -> PEP_cyt  + Pi_str ~

TX_G6P_cyt:
G6P_str  + Pi_cyt -> G6P_cyt + Pi_str ~

TX_DHAP_cyt:
DHAP_str + Pi_cyt -> DHAP_cyt + Pi_str ~

TX_GAP_cyt:
GAP_str  + Pi_cyt -> GAP_cyt + Pi_str ~

TX_PGA_cyt:
PGA_str  + Pi_cyt -> PGA_cyt + Pi_str ~

TX_MAL_cyt:
MAL_cyt  <> MAL_str ~

TX_OAA_cyt:
OAA_cyt <> OAA_str ~
```

Figure C.1 – Model of the glycolytic reactions in chloroplast in *.spy* format, as described in Section 2.2.2 and illustrated in Figure 2.5. See List of Abbreviations for metabolite and reaction abbreviations. '_cyt' and '_str' indicate metabolites localised in cytosol and stroma, respectively.

APPENDIX D

Model of the mitochondrial metabolism

```
########################################################################
###################   MITOCHONDRIAL METABOLISM   ###################
########################################################################

Structural()

External(MAL_cyt, PYR_cyt, OAA_cyt, Proton_ims)
External(NADHWork_mit, ATPwork_mit)

########################################################################
################### REACTIONS OF THE TCA CYCLE ##########################
########################################################################

PDH_mit:
PYR_mit + NAD_mit + CoASH_mit + Proton_mit -> ACoA_mit + NADH_mit ~

CITSynth_mit:
OAA_mit + ACoA_mit <> CIT_mit + CoASH_mit ~

ACN_mit:
CIT_mit <> IsoCIT_mit ~

IDH_mit:
IsoCIT_mit + NAD_mit + Proton_mit -> AKG_mit + NADH_mit ~

AKGDH_mit:
AKG_mit + CoASH_mit + NAD_mit + Proton_mit -> SCoA_mit + NADH_mit ~

SCS_mit:
SCoA_mit + ADP_mit + Pi_mit <> SUC_mit + ATP_mit + CoASH_mit ~

SDH_mit:
SUC_mit + Q_mit <>  Fumerate_mit + QH2_mit ~

FUM_mit:
Fumerate_mit <> MAL_mit ~

NADMDH_mit:
MAL_mit + NAD_mit + Proton_mit <> OAA_mit + NADH_mit ~

########################################################################
#######################   TRANSPORT REACTIONS   ########################
########################################################################

TX_MAL_mit:
MAL_mit <>  MAL_cyt ~

TX_OAA_mit:
OAA_cyt <>  OAA_mit ~
```

```
TX_PYR_mit:
PYR_mit <>  PYR_cyt ~

#####################################################################
############ MITOCHONDRIAL ELECTRON TRANSPORT CHAIN ################
#####################################################################

Complex_I:
NADH_mit + Q_mit + Proton_mit -> NAD_mit + QH2_mit + Proton_ims
+ NADHWork_mit ~

Complex_III:
QH2_mit +  Cyt_ox_mit +  Proton_mit -> Q_mit + Cyt_red_mit + Proton_ims ~

Complex_IV:
Cyt_red_mit + Proton_mit -> Cyt_ox_mit + Proton_ims ~

Complex_V:
3ADP_mit + 3Pi_mit + 12Proton_ims -> 12Proton_mit +  3ATP_mit ~

ATPSink_mit:
ATP_mit -> ADP_mit + Pi_mit + ATPwork_mit ~
```

Figure D.1 – Model of mitochondrial metabolism containing reactions of TCA cycle and mitochondrial ETC in *.spy* format, as described in Section 2.2.3 and illustrated in Figures 2.6 and 2.7. See List of Abbreviations for metabolite and reaction abbreviations. '_cyt', '_mit' and '_ims' indicate metabolites localised in cytosol, mitochondria and intermembrane space, respectively.

APPENDIX E

AraCyc identifiers of the reactions in the extended model

Abbreviation	AraCyc Identifier
ACN	ACONITATEDEHYDR-RXN
AKGDH	2OXOGLUTDECARB-RXN
Ald1	F16ALDOLASE-RXN
Ald2	SEDOBISALDOL-RXN
CITSynth	CITSYN-RXN
Complex I	NADH-DEHYDROG-RXN
Complex III	1.10.2.2-RXN
Complex IV	CYTOCHROME-C-OXIDASE-RXN
ENO	2PGADEHYDRAT-RXN
FBPase	F16BDEPHOS-RXN
FUM	FUMHYDR-RXN
GAPDH	1.2.1.13-RXN
GAPDHP	GAPOXNPHOSPHN-RXN
GlyM	3PGAREARR-RXN
IDH	ISOCITDEH-RXN
NAD-MDH	MALATE-DEH-RXN
NADP-MDH	MALATE-DEH-RXN
NDPK	NUCLEOSIDE-DIP-KIN-RXN
PDH	RXN0-1134
PEPC	PEPCARBOXYKIN-RXN
PFK	6PFRUCTPHOS-RXN
PFP	2.7.1.90-RXN
PGI	PGLUCISOM-RXN
PGK	PHOSGLYPHOS-RXN
PGM	PHOSPHOGLUCMUT-RXN
PK	PEPDEPHOS-RXN
R5Piso	RIB5PISOM-RXN
Ru5PK	PHOSPHORIBULOKINASE-RXN
Rubisco	RIBULOSE-BISPHOSPHATE-CARBOXYLASE-RXN
SBPase	SEDOHEPTULOSE-BISPHOSPHATASE-RXN
SCS	SUCCCOASYN-RXN
SDH	SUCCINATE-DEHYDROGENASE-(UBIQUINONE)-RXN
StPase	RXN-1826
StSynth	GLYCOGENSYN-RXN
SuSyn	SUCROSE-SYNTHASE-RXN
TKL1	1TRANSKETO-RXN
TKL2	2TRANSKETO-RXN
TPI	TRIOSEPISOMERIZATION-RXN
UGPase	GLUC1PURIDYLTRANS-RXN
X5Piso	RIBULP3EPIM-RXN

APPENDIX F

UML representation of the ScrumPy add-on used for integrating metabolic models with gene expression data

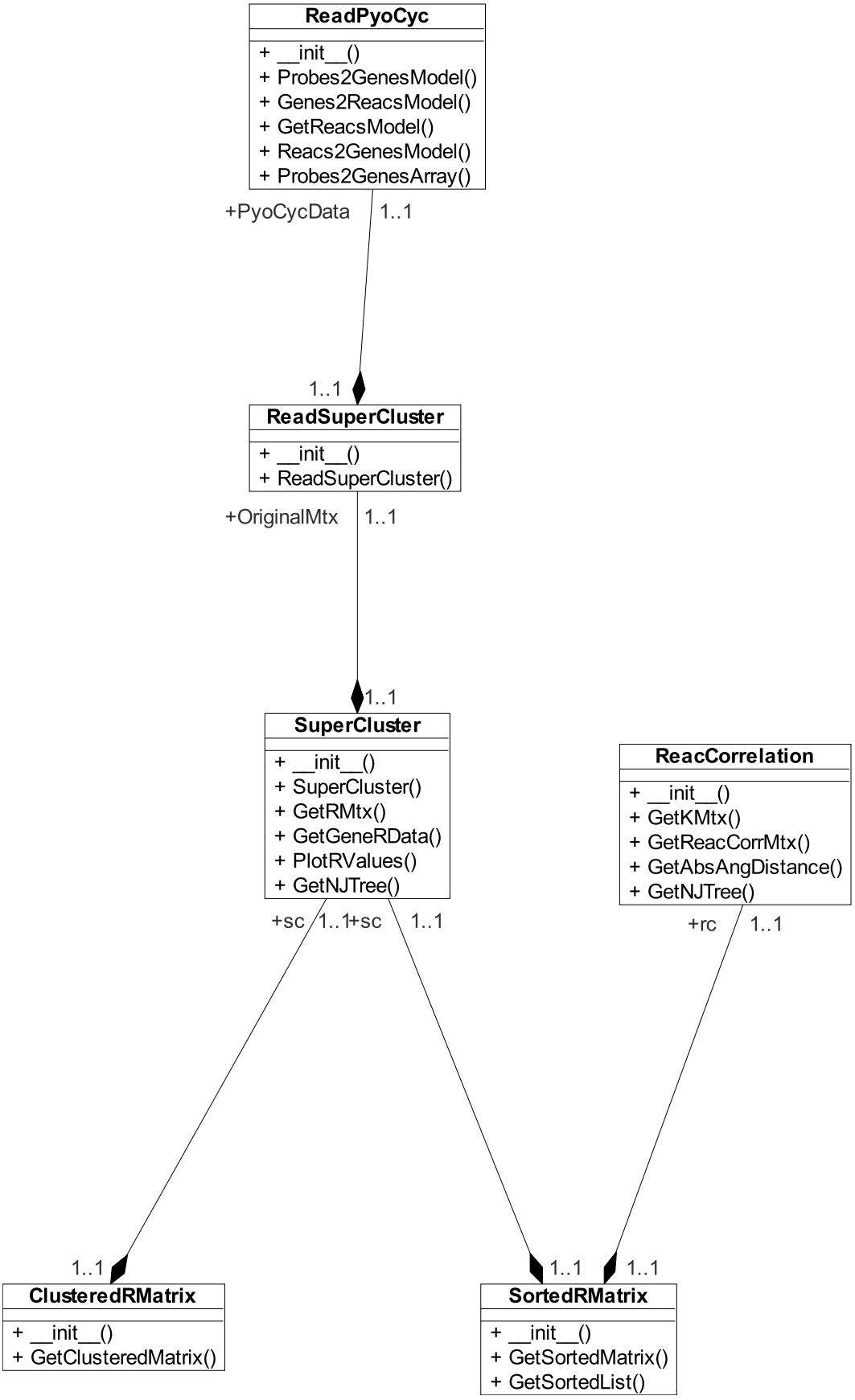


Figure F.1 – UML representation of the ScrumPy add-on used for integrating metabolic models with gene expression data

APPENDIX G

Contents of the CD

Table G.1 – Files and directories included in the CD accompanying this thesis.

File/Directory name	Contents
chlorocytomito_rmtx	Heatmap showing Pearson's correlation matrix generated from the expression matrix representing the correlation between genes coding for reactions in the combined model.
chlorocytomito_exprFullTree.pdf	Dendrogram representing the correlation between the expression profiles of genes coding for reactions in the model.
chlorocytomito_clusteredrmtx.pdf	Heatmap showing Pearson's correlation matrix representing the correlation between the expression profiles of genes coding for reactions in the combined model. Rows and columns of this matrix were hierarchically clustered.
chlorocytomito_srmtx.pdf	Heatmap showing the Pearson's correlation matrix whose rows and columns were hierarchically clustered based on reaction correlation coefficients (RCCs) representing correlation between fluxes carried by reactions in the model and Pearson's correlation coefficients representing correlation between the expression profiles of genes coding for reactions in the model, respectively.
cluster_a.pdf	Heatmap representing within- and cross- pathway correlations. It also shows the correspondence between the correlation profiles of rubisco (highlighted with with a box around the gene name) and other genes coding for reactions in chloroplast and cytosol.
cluster_1.pdf	Corresponding correlation profiles of the genes coding for reactions in Cluster 1 that have correlated flux. Gene names are suffixed with reaction names and details about the compartment in the metabolic model from which it was extracted.
isoforms.pdf	Heatmap representing the correlation profiles of genes coding for isoforms of pyruvate kinase (PK) in different compartments.
compartment_predictions.txt	Tab delimited text file containing the localisation of genes coding for reactions in the integrated model predicted using the approach described in Chapter 5.
supercluster.txt	Tab delimited text file containing expression data of over 225000 genes in approximately 60 experiments. Data obtained from NASCArrays (March, 2010).
python_scripts	Contains the Python scripts used for integrating RCCs with gene expression profiles. The scripts require ScrumPy and PyoCyc to be installed on a Linux platform. The UML diagram for the scripts is available in Appendix F.

APPENDIX H

Publication

[1] A. B. Chokkathukalam, M. G. Poolman, C. Ferrazzi, and D. A. Fell, "Expression profiles of metabolic models to predict the compartmentation of enzymes in multi-compartmental systems," in Proceedings of the German Conference on Bioinformatics (GCB 2009), Halle (Saale) , Germany, September 28-30, 2009 (I. Grosse, S. Neumann, S. Posch, F. Schreiber, and P. F. Stadler, eds.), vol. 157 of LNI, GI, 2009.

Expression profiles of metabolic models to predict compartmentation of enzymes in multi-compartmental systems

Achuthanunni Chokkathukalam, Mark Poolman, Chiara Ferrazzi and David Fell

cbaunni@brookes.ac.uk

Abstract: Enzymes and other proteins coded by nuclear genes are targeted towards various compartments in the plant cell. Here, we describe a method by which localisation of enzymes in a plant cell may be predicted based on their transcription profile in conjunction with analysis of the structure of the metabolic network. This method uses reaction correlation coefficients to identify reactions in a metabolic model that carry similar flux.

First a correlation matrix for the expression of genes of interest is calculated and the columns clustered hierarchically using the correlation coefficient. The rows clustered using reaction correlation coefficients. In the resulting matrix, we show that the genes in a particular compartment are clustered together and compartmental predictions, with respect to a reference gene can be readily made.

1 Introduction

Spatial organisation of metabolism and other cellular functions is a well known feature of plant cells. Enzymes and other proteins coded by nuclear genes are targeted towards various compartments in the plant cell with the help of the targeting information within their amino acid sequence. Identifying the localisation of proteins is thus an important step towards a broader understanding of the cellular function as a whole and may help in determining the role of thousands of uncharacterised proteins predicted by the genome sequencing projects. Modern organelle-focused experimental approaches can identify proteins in a given compartment. However, reliable protein localisation requires that the technique used must be able to distinguish between genuine organelle residents and contaminating proteins [DDWL04]. Although reasonably pure preparations of some organelles can be achieved, there are many difficulties associated with measuring and characterising proteins that are in a compartment [DHS$^+$06]. Nevertheless, a variety of experimental methods are currently being used to identify protein localisation. Recently chimeric fusion proteins (FPs) and mass spectrometry (MS) techniques have been successfully employed to deduce the localisation of approximately 1100 and 2600 proteins, respectively [HVTF$^+$06]. Although these techniques have accelerated the flow of protein localisation information, the subcellular location of the majority of proteins in a plant cell is still not known.

A relatively simple, low-cost and rapid means to tackle this issue is to employ bioinfor-

matic targeting algorithms to predict protein localisation from amino acid sequence. A number of software tools exists, including TargetP [EBvHN07], Predotar [SPLL04], iPSORT [BTM+02], SubLoc [HS01], MitoProt II [CV96], MITOPRED [GFS04], PeroxiP [EEvHC03], and WoLF PSORT [HPO+07], which can predict proteins targeted towards plastid, cytosol, nucleus, mitochondria, peroxisome or the endoplasmic reticulum. However, the output of such programs has been found to be somewhat inconsistent with each other, or with experimentally determined results [HVTFM05], making them unreliable for some analyses.

The advent of whole-system approaches such as microarrays and metabolomics and the accumulation of such high-throughput data have created new opportunities for studying how reactions are coordinated to meet cellular demands. Microarray experiments monitor the expression of thousands of genes simultaneously. Grouping together genes of similar expression pattern is a general starting point in the analysis of expression data. Similarity between genes is measured by the correlation of their expression profiles and hierarchical clustering methods are used to partition data into clusters of genes exhibiting similar expression patters [IBB04]. Numerous studies have shown that co-expression patterns of gene expression across many microarray datasets form modules of genes that are functionally correlated [WPM+06, MDO+08]. Recently this approach was successfully employed in identifying new genes involved in cellulose synthesis in plants [PWM+05].

Here, we describe a method by which localisation of enzymes may be predicted based on the co-expression profiles of genes coding for reactions in a structural model of plant carbon metabolism. Structural models contain stoichiometries of reactions in a metabolic system. Based on the correlation between these reactions, it can be represented hierarchically as a metabolic tree in which the root node represents the complete system, leaf nodes represent individual reactions, and the intermediate nodes represent metabolic modules capable of the net interconversion of metabolites common to reactions inside and outside the module [PSPF07]. Our technique uses reaction correlation profiles generated from metabolic models together with expression correlation profiles obtained from the microarray data to identify the distribution of enzymes in a particular compartment with respect to the experimentally determined location of a protein representing that compartment.

2 Materials and methods

2.1 Construction of the model of plant carbon metabolism

A structural model of plant carbon metabolism including plastid and cytosol compartments was constructed (Figure 1). The model contains reactions of the Calvin cycle, light reactions and glycolysis and is based, in part, on previous models of plant metabolism constructed in our group [PFR03, Ass05]. Protons, CO_2, pyruvate and sucrose were made external (metabolites that are in constant exchange with the extracellular environment) yielding a model with a total of 53 reactions and 49 metabolites. Reversibility of the reactions was determined based on literature. All modelling and model analysis were performed

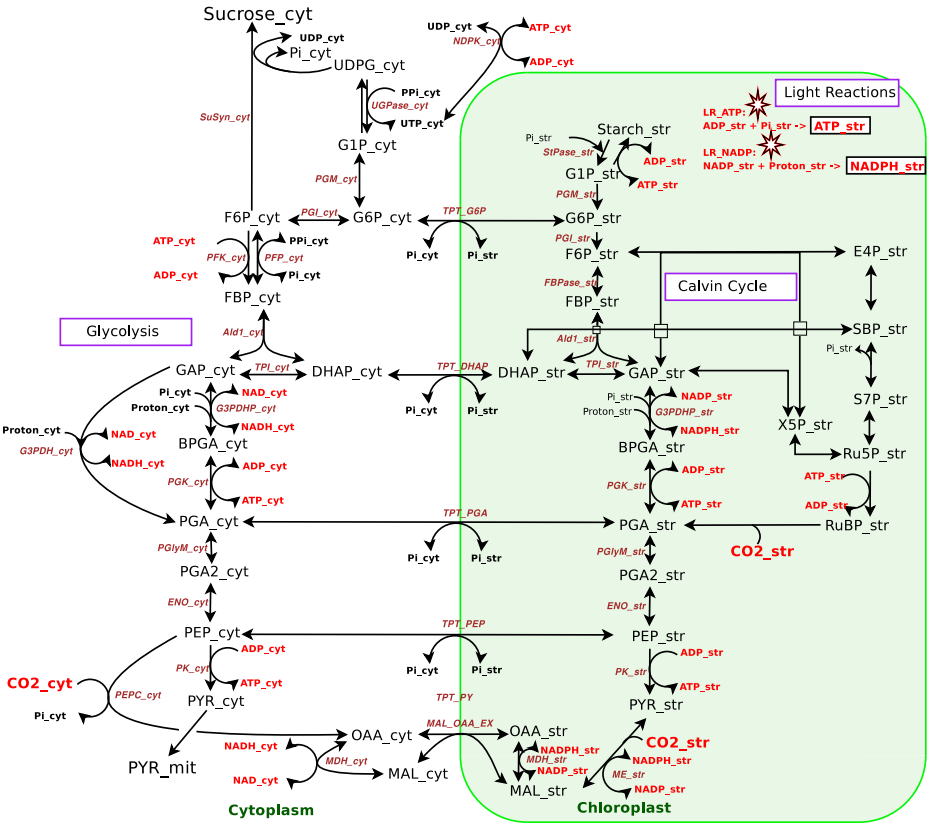


Figure 1: Reaction schema of the model of plant carbon metabolism. For simplicity, the light reactions are depicted here as two separate reactions producing ATP and NADPH. Protons, CO_2 and sucrose are considered external. '_str' and '_cyt' represent the compartments stroma and cytosol, respectively. Notice the transporters connecting reactions of the plastid and the cytosol.

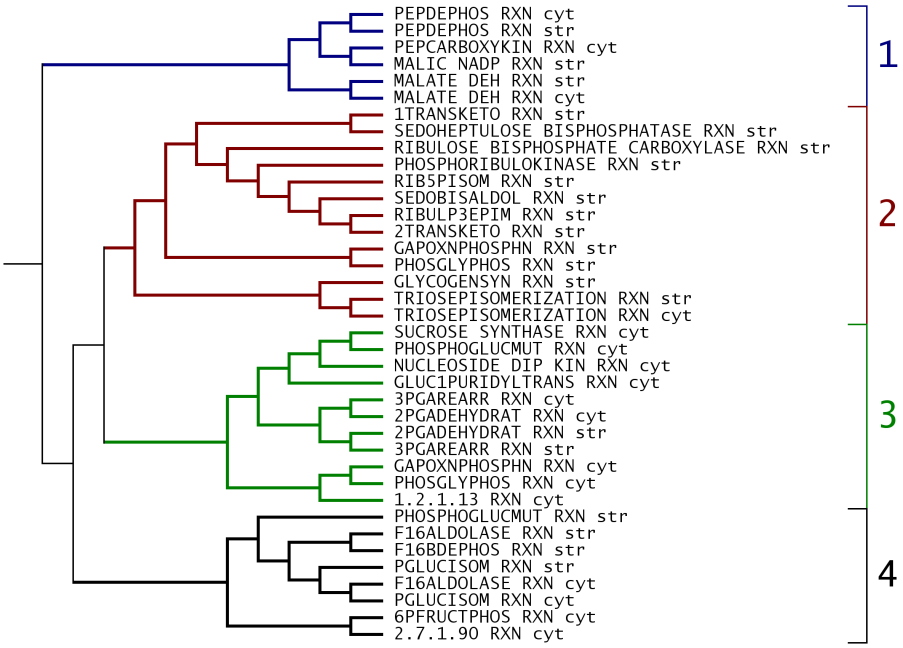


Figure 2: Metabolic tree constructed from the model showing four separate clusters containing reactions capable of net interconversion of metabolites; A. Reactions of the Malate/Oxaloacetate shuttle. B. Calvin cycle reactions. C. Reactions of glycolysis. D. Reactions involved in the regeneration of cytosolic UDP. '_str' and '_cyt' represent the compartments chloroplast and cytosol, respectively.

using the metabolic modelling tool ScrumPy (http://mudshark.brookes.ac.uk) [Poo06].

The model represents the formation of sucrose and pyruvate from the Calvin cycle intermediates transported to the cytosol via specific transport proteins. It contains several reactions such as phosphoglyceromutase, enolase, pyruvate kinase and malate dehydrogenase that are active in both the chloroplast and cytosol. Presence of these reactions in the model will enable us not only to identify their distribution between the compartments but also to distinguish isoforms of genes that code for same reactions in both the compartments. This model is publically available as SBML or in the ScrumPy '.spy' format (http://mudshark.brookes.ac.uk/index.php/User:Cbaunni).

2.2 Expression data analysis of genes coding for reactions in the model

The gene to reaction associations describe the dependence of reactions on genes. The gene to reaction associations in the model were mapped using the AraCyc [ZFT+05] database (http://www.arabidopsis.org/biocyc/index.jsp). The result is a set of genes that potentially code for all the reactions in the model.

The expression data for analysing these genes were obtained from the Nottingham *Arabidopsis* Stock Centre's (NASC) microarray database (http://affymetrix.arabidopsis.info/). The 'super bulk gene' file containing nearly 3500 hybridisations, each with expression level measurements for over 22000 genes represented on the ATH1 array was downloaded (http://affymetrix.arabidopsis.info/narrays/help/usefulfiles.html, March 2009). Expression data from individual experiments were log-transformed; no further modification or scaling was made on the data unless otherwise specified. All microarray data analysis was performed using custom modules designed for ScrumPy.

Expression data for genes ultimately coding for reactions in the model were extracted and a large-scale correlation analysis of expression values between these genes were performed essentially as described by Causton *et al.* [CQB03] by calculating the Pearson's correlation coefficient.

2.3 Clustering and analysis of the correlation matrix

A metabolic tree was generated from the model using the method described in [PSPF07] (Figure 2). The order of the reactions in this tree was used to sort the genes along the rows of the correlation matrix.

The columns of the matrix were hierarchically clustered based on the Pearson's correlation coefficient and an expression correlation tree was generated (Figure 3). Leaves of this tree represent genes in the model and the intermediate nodes are clusters that represent genes sharing similar functions. The columns of the correlation matrix were then sorted in the order of the leaves of the expression correlation tree.

The correlation matrix was imported into TM4-MeV (http://www.tm4.org/mev.html) for visualisation as heatmap [ESBB98]. The metabolic trees were visualised using MEGA phylogenetic tree editor (http://www.megasoftware.net/) [KNDT08].

3 Results and Discussion

3.1 Identification of correlated genes sharing similar flux

Metabolic tree generated from the model contain four separate clusters, each representing reactions capable of net interconversion of metabolites (Figure 2). It is notable that reactions of the Calvin cycle and glycolysis are represented as separate nodes on the tree. Clustering the rows of the correlation matrix based on the genes coding for reactions represented in these nodes can rearrange the heatmap vertically based on the similarities in flux. On the other hand, hierarchically clustering the columns of the correlation matrix grouped genes horizontally depending on their levels of expression. Doing so resulted in the formation of clusters in the heatmap representing genes that are expressed together and code for enzymes that share a similar flux (Figure 4).

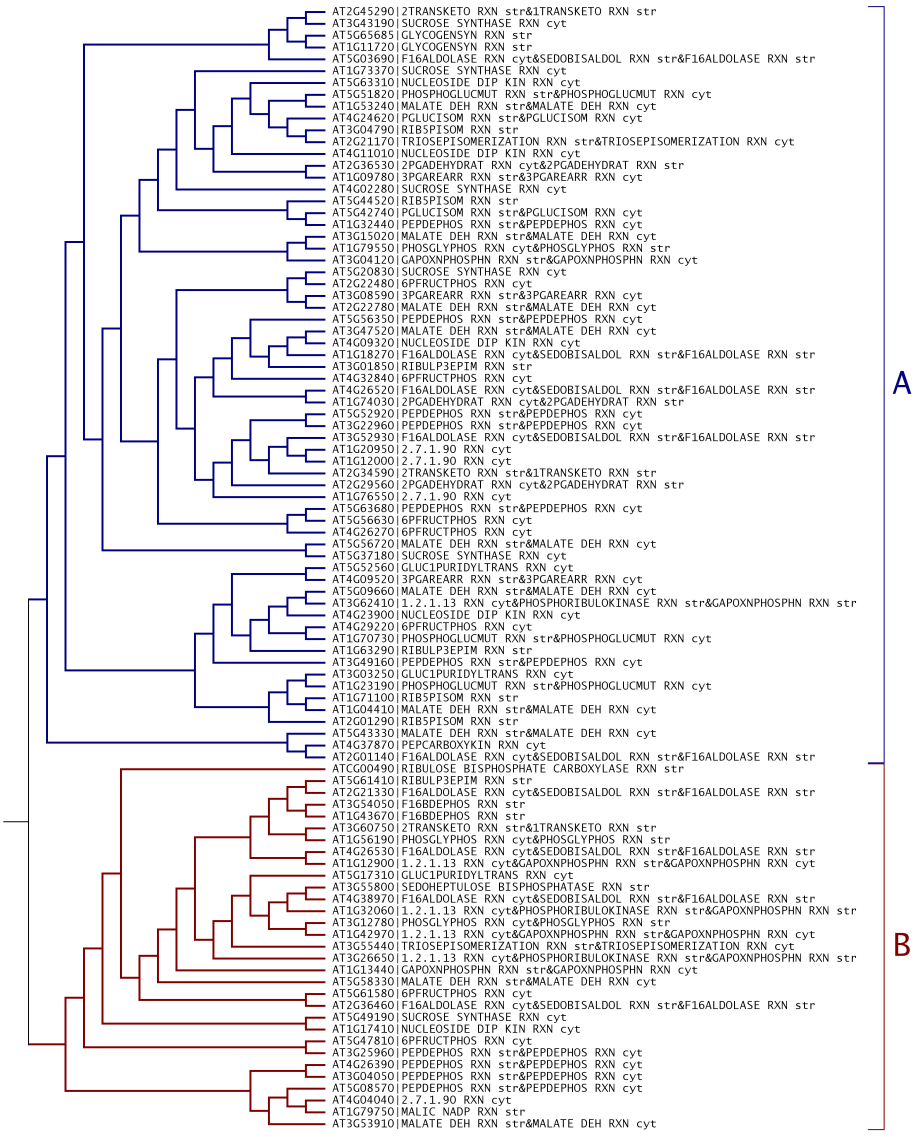


Figure 3: Expression correlation tree generated by hierarchically clustering correlation coefficients of genes coding for reactions in the model showing two separate clusters. A. Genes that predominantly code for reactions in the cytosol correlate with each other B. Genes coding for Calvin cycle intermediates cluster together. '_' is used to separate genes from reactions and '&' is used to distinguish reactions that the gene code for. '_str' and '_cyt' represent the compartments chloroplast and cytosol, respectively.

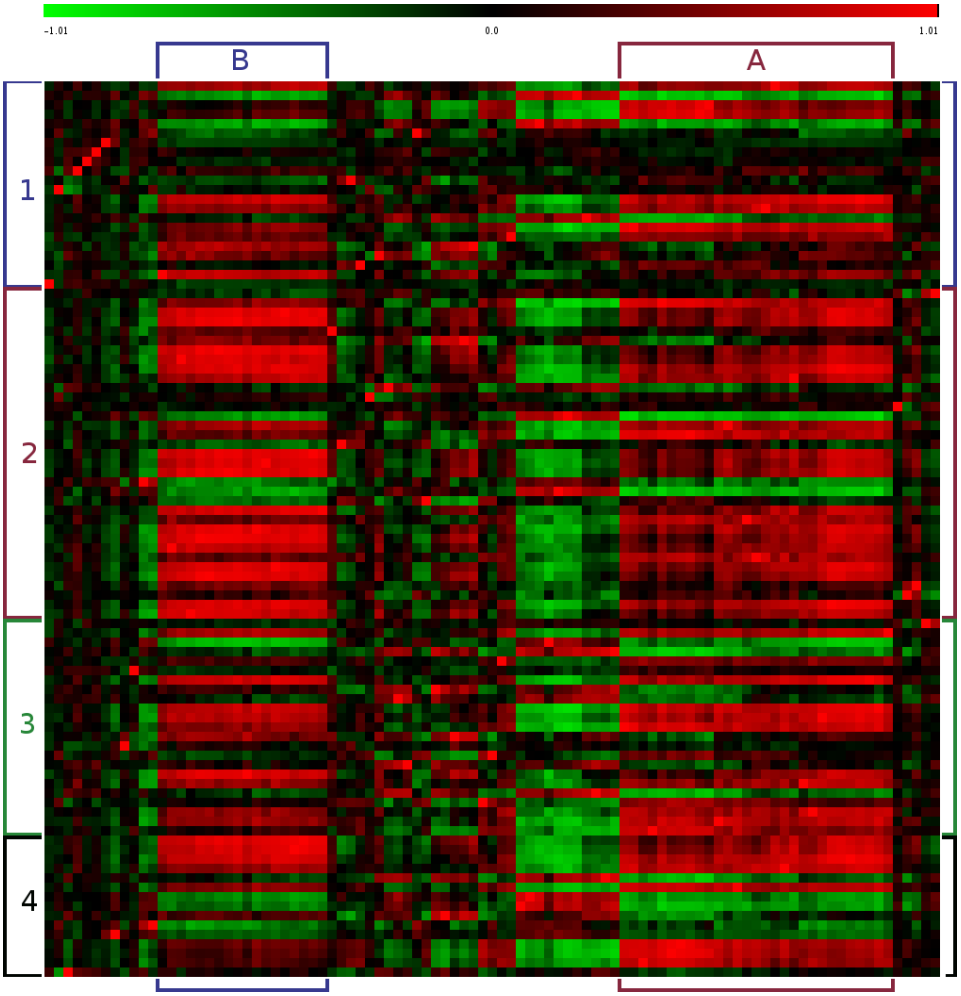


Figure 4: Correlation matrix generated from the expression values of genes coding for reactions in the steady state model. The correlation coefficient ranges from -1 (green) for perfect anticorrelation to +1 (red) for perfect correlation, with zero (black) indicating no relationship. Columns were sorted based on the clustering expression correlation coefficient and rows sorted by clustering based on reaction correlation coefficient. 'A' and 'B' represent two distinct clusters observed in the correlation matrix (Figure 3). Correlated genes in cluster 'A' were found to be highly correlated with reference genes known to be localised in the chloroplast. Whereas correlated genes in cluster 'B' showed higher correlation with genes localised in the cytoplasm. 1, 2, 3 and 4 represent clusters in the metabolic tree representing reactions capable of net interconversion of metabolites (Figure 2).

We found that genes coding for reactions in the Calvin cycle are found to be tightly correlated between each other and they cluster together. The same holds true for genes coding for glycolysis reactions. Isoforms of some Calvin cycle genes anticorrelate with other genes coding for reactions of the Calvin cycle. However, those genes that were anticorrelated with the genes of Calvin cycle reactions are found to be tightly correlated with genes of the glycolysis reactions, and vice versa. Similar cases can also be observed in case of the isoforms of glycolytic genes.

A previous study on the transcriptional coordination of metabolic network in *Arabidopsis* suggested that genes coding for reactions in a pathway show tighter levels of correlation [WPM+06]. Results from our study correlates with the above observation and also suggests that the expression profiles of genes can be used to distinguish their compartmentation.

3.2 Identifying compartmentation of genes

Though, this technique is efficient in clustering genes based on their compartmentation, identification of the compartment itself requires a reference gene whose localisation is already known. For example, the plastidic ribulose biphosphate carboxylase (Rubisco) gene ATCG00490 was used as the reference to identify genes localised in the chloroplast. Compartments are identified by filtering out genes that are highly correlated with the reference gene.

The results were compared with the various bioinformatic tools described in Section 1. Comparison with predictions made by bioinformatic tools as a whole was not possible as many of these tools were directed towards particular compartments. Compartmentation of genes that were predicted to be in the chloroplast showed good agreement with tools such as TargetP and Predotar, whereas mitochondrial predictions correlated with MITOPRED and MitoProt II predictions.

This approach was used to predict the localisation of the complete set of genes coding for the reactions in a model containing reactions of the chloroplast, cytosol and mitochondria. Given a good quality microarray expression data containing sufficient experiments that allow reliable statistical analysis, this technique can be used more generically. With the large number of publically available metabolic networks and expression data, this approach may significantly contribute to the identification of enzyme localisation in many different eukaryotic systems.

References

[Ass05] H. Assmus. *Modelling Carbohydrate Metabolism in Potato Tuber Cell*. PhD thesis, Oxford Brookes University, 2005.

[BTM+02] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18(2):298–305, 2002.

[CQB03] Helen C. Causton, John Quackenbush, and Alvis Brazma. *Microarray gene expression data analysis: a beginner's guide*. Wiley-Blackwell, 2003.

[CV96] M.G. Claros and P. Vincens. Computational Method to Predict Mitochondrially Imported Proteins and their Targeting Sequences. *European Journal of Biochemistry*, 241:779–786, 1996.

[DDWL04] T.P.J. Dunkley, P. Dupree, R.B. Watson, and K.S. Lilley. The use of isotope-coded affinity tags (ICAT) to study organelle proteomes in Arabidopsis thaliana. *Biochem. Soc. Trans.*, 32(3):520–523, 2004.

[DHS+06] T.P.J Dunkley, S. Hester, I.P. Shadforth, J. Runions, T. Weimar, S.L. Hanton, J.L. Griffin, C. Bessant, F. Brandizzi, C. Hawes, R.B Watson, P. Dupree, and K.S. Lilley. Mapping the Arabidopsis organelle proteome. *Proc. Natl. Acac. Sci. USA*, 103(17):6518–6523, 2006.

[EBvHN07] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen. Locating proteins in the cell using TargetP, SignalP, and related tools. *Nature Protocols*, 2:953–971, 2007.

[EEvHC03] O. Emanuelsson, A. Elofsson, G. von Heijne, and S. Cristobal. In Silico Prediction of the Peroxisomal Proteome in Fungi, Plants and Animals. *Journal of Molecular Biology*, 330:443–456, 2003.

[ESBB98] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868., 1998.

[GFS04] C. Guda, E. Fahy, and S. Subramaniam. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, 20(11):1785–1794, 2004.

[HPO+07] P. Horton, K-J. Park, T. Obayashi, N. Fujita, H. Harada, C.J. Adams-Collier, and K. Nakai. WoLF PSORT: Protein Localization Predictor. *Nucleic Acids Research*, pages 1–3, 2007.

[HS01] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.

[HVTF+06] J.L. Heazlewood, R.E. Verboom, J. Tonti-Filippini, I. Small, and A.H. Millar. SUBA: The Arabidopsis Subcellular Database. *Nucleic Acids Research*, 00:1–6, 2006.

[HVTFM05] J.L. Heazlewood, R.E. Verboom, J. Tonti-Filippini, and A.H. Millar. Combining experimental and predicted datasets for determination of the subcellular location of proteins in Arabidopsis. *Plant Physiol.*, 139(2):598–609, 2005.

[IBB04] J. Ihmels, S. Bergmann, and N. Barkai. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003, 2004.

[KNDT08] S. Kumar, M. Nei, J. Dudley, and K. Tamura. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics*, 9(4):299–306, 2008.

[MDO+08] M. Menges, R. Doczi, L. Okresz, P. Morandini, L. Mizzi, M. Soloviev, J.A.H. Murray, and L. Bogre. Comprehensive gene expression atlas for the Arabidopsis MAP kinase signalling pathways. *New Phytologist*, 179(3):643–662, 2008.

[PFR03] M.G. Poolman, D.A. Fell, and C.A. Raines. Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. *Eur. J. Biochem*, 270:430–439, 2003.

[Poo06] M.G. Poolman. ScrumPy - metabolic modelling with Python. *IEE Proceedings Systems Biology*, 153(5):375–378, 2006.

[PSPF07] M.G. Poolman, C. Sebu, M.K. Pidcock, and D.A. Fell. Modular decomposition of metabolic systems via null-space analysis. *Journal of Theoretical Biology*, 249(4):691–705, 2007.

[PWM+05] S. Persson, H. Wei, J. Milne, G.P. Page, and C.R. Somerville. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc Natl Acad Sci USA*, 102:8633–8638, 2005.

[SPLL04] I. Small, N. Peeters, F. Legeai, and C. Lurin. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, 4(6):1581–90, 2004.

[WPM+06] H. Wei, S. Persson, T. Mehta, V. Srinivasasainagendra, L. Chen, G.P. Page, C. Somerville, and A. Loraine. Transcriptional Coordination of the Metabolic Network in Arabidopsis. *Plant Physiol.*, 142(2):762–774, 2006.

[ZFT+05] P. Zhang, H. Foerster, C.P. Tissier, L. Mueller, S. Paley, P.D. Karp, and S.Y. Rhee. MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.*, 138:27–37, 2005.