NTNU
Norwegian University of
Science and Technology

Genome-scale constraint-based metabolic modeling and analysis of Nannochloropsis Sp.

Nhung Pham

Biotechnology
Submission date: May 2016
Supervisor: Martin Frank Hohmann-Marriott, IBT

Norwegian University of Science and Technology
Department of Biotechnology

Abstract

Genome scale metabolic models provide a link between genomic and metabolic information. The constraint-based approach is increasingly important for quantitatively analyzing the essential aspects of metabolic pathways, for example finding the reaction fluxes. This thesis was aimed at constructing a genome scale metabolic model for *Nannochloropsis*, a unicellular marine alga that offers a potential feedstock for biofuel production due to its high lipid content. The model reconstruction was carried out following a procedure outlined by Thiele, I. and Palsson B.Ø. in 2010 [1]. After obtaining an automatically generated model using the PlantSeed platform, this draft reconstruction was manually curated. The final model consists of 383 genes cover 4.8% of the total gene in the genome. A total of 987 reactions and 1024 metabolites are included in the model and were assigned to biochemical pathways. All gaps related to the synthesis of biomass precursor were eliminated. At this stage the model can synthesize 70 out of 89 biomass precursors. Since *Nannochloropsis* is interesting as a possible producer of lipids to be utilized in biodiesel production, the synthesis of triacylglycerol (TAGs) had a special focus in this project. The TAG synthesis was reconstructed as accurately as possible. Using this model, TAGs is produced at the flux rate of 23.0138. The generated model shares some common characteristics with models of other photosynthetic organisms. These common characteristics include network topology, distribution of reactions and metabolites in compartments and pathways. The thesis provides the first genome-scale metabolic model for *Nannochloropsis*, and is a starting point for future improvement and validation for an accurate simulation of lipid and biomass production.

Acknowledgment

This work would not have been completed without the participation and assistance of many people. First of all, I would like to express my gratefulness to my supervisors, Assoc. Prof. Martin F. Hohmann-Marriott and Prof. Eivind Almaas for suggesting this study and offering me the freedom to explore so widely and independently the field of systems biology. They have been available to me and provide essential guidance and suggestion whenever needed.

Secondly, I would like to especially thank Gunvor Røkke for all of her help and supporting. More than an advisor, she is also a friend, a sister and a companion who always available whenever I was depress and stressful. I am really grateful for her patient to help me with this project. In fact, the reconstruction of triacylglycerol synthesis was not able to complete without her. In addition, she was the one who support my first steps into the field of systems biology. Before doing this project, I have no knowledge about MATLAB. But thanks to a crash course offered by Gunvor, my study flew smoothly. I still remember those first days struggling with writing script, a simple script with few lines could take me hours. Yet, eventually, at the moment I can write many scripts to support the construction of the model in a blink of an eye (alright, I was too exaggerated, but, writing a script is not too difficult for me as before).

Despite whoever you are and whatever you do, you will not able to finish your work without supporting from your beloved ones. So I would like to thank my family for all of their carefulness and worries during the time I carried out this project.

Finally, I would like to thanks all of my friends who also did their master projects at the same time with me. Their motivations had inspired me a lot to overcome this important duration.

Content

List of figures

List of tables

Abbreviations

BRENDA	Braunschweig Enzyme Database
CBMS	Constraint- based models
CMR	Comprehensive Microbial Resource
COBRA	COnstraint-Based Reconstruction and Analysis framework
CYGD	Comprehensive Yeast Genome Database
EC	Enzyme Commission
FBA	Flux balance analysis
GA3P	Glyceraldehyde 3 phosphate
GNS	Genome-scale
GPR	Genes-Proteins-Reactions Association
IMG	Intergrated Microbial Genome
KEGG	Kyoto Encylopedia of Genes and Genomes
LP	Linear Programming
MEP/DOXP	The non-mevalonate pathway
MILP	Mixed Integer Linear Programming
MIQP	Mixed-Integer Quadratic Programming
MVA	the mevalonate pathway
ORF	Open reading frame
PRPP	5-phosphoribosyl-1-pyrophosphate
RNCs	Root Non Consumed metabolites
RNPs	Root Non Produced metabolites
SGD	*Saccharomyces* Genome Database
TAGs	Triacylglycerol
TC	Transport Commission
TCA cycle	The citric acid cycle

1 Introduction

The development of genome sequencing technology has provided a powerful tool to investigate the interaction between genotypes and phenotypes of cells. Genome sequencing technology has grown to employ millions scientists worldwide in the passing years [2]. Since the genome of the bacterium *Haemophilus influenzae* Rd was sequenced as the first one in 1995 [3] , genomic sequencing techniques have improved significantly in term of quality and quantity. After the emergence of the automatic Sanger method, often referred as the first-generation technology, the next generation sequencing technique allows the analysis of a large number of organisms by much faster generation of genomes with lower costs [4]. The cost to obtain a genome sequence has dropped sharply from $1000.000.000 in 2001 to only $1000 in 2015 [5]. Nowadays, almost every organism can be sequenced; widen the availability for scientists to get insight of the organism's genomes.

With more than 16000 genomes from organisms of all domains available on NCBI [6], together with the advance of high -throughput technologies in many 'omics' fields including transcriptomic, proteomic and metabolomics, research on putting all this genetic information into a system, and using it to model metabolic networks has become more and more common [7] [8]. Because of the abundance of genetic information, the metabolic events inside a cell can be now thoroughly studied, and the single metabolic reactions can be used to generate a bigger picture. Beyond research about individual components in an organism, the question in the current post-genome era is how to use the huge amount of high- throughput biological data to understand and to predict the function of cells as a whole [8].

Predicting the behavior of living organisms from their individual components is difficult due to the dynamic and complex natures of living systems. Therefore, systems biology is currently considered a new and better alternative for studying cellular behavior [9], instead of the traditional approach, that studies details of the individual components but does not focus on the systematic interaction between the biological and environmental factors that control the cellular phenotypes. Systems biology serves as a new lens for looking at the interactions of systems consisting of biological components such as molecules, cells, organisms or entire species as a whole [9]. Systems biology, involving modeling and analysis of metabolic pathways, but also

regulatory and signal transduction network, has grown rapidly as a research field, and has become central in all areas of biology and medicine [10] [8].

Among computational approaches in System biology, constraint based modelling is used extensively for simulation of network behavior, because it does not require kinetic data [11]. This is an advantage, as kinetic data can be hard to collect for all components making up a network. The constraint based modelling approach involves reconstruction and analysis of genome scale metabolic networks, based on data of metabolites and their reactions [10] [8] . Constraint-based models (CBMs) have mainly been used for metabolic networks, but it may as well be used for signaling networks, transcriptional regulation networks and macromolecule synthesis [11]. At genome-scale level, CBMs are shown to be useful for many applications, such as guidance of metabolic engineering, hypothesis generation, and research about the relationships of multi-species and network property [8].

Recently, a large number of CBMs has been published for many organisms both among bacteria, archaea and eukaryotic cells [8]. Out of these three domains, bacteria have gotten the most attention. In contrast, the number of CBMs for plants and algae is still relatively small [8]. Up to now, CBMs have only been built for *Arabidopsis thaliana* and *Chlamydomonas reinhardtii* [8] [12] [13] [14]. Thus, constraint-based modelling of the metabolism of plants and algae has great possibilities for future researchers.

Algae are also currently in the spotlight because of the need for developing a new source for sustainable energy. This is because many algae produce relatively large amounts of fatty acids, which could be extracted and converted to biodiesel [15] [16]. The heterokont microalga *Nannochloropsis* is considered one of the most promising organisms for alternative biofuel production due to its abilities to produce and to store lipid as much as 60 % of its own weight in certain conditions [17] [18, 19]. In addition, *Nannochloropsis* is also well-known for its nutritional value and its ability to produce pigments like zeaxanthin and astaxanthin, and polyunsaturated fatty acids (for example eicosapentaenoic acid, more commonly known as EPA) [20]. Thanks to the availability of multiple sequenced *Nannochloropsis* genomes [21] [18, 22, 23], it is feasible to develop genome-scale reconstruction of *Nannochloropsis*. The CBMs can provide a comprehensive understanding of its metabolism, as well as being useful when planning metabolic engineering of this organism.

The objective of this master project is to build a genome-scale reconstruction of the metabolism in *Nannochloropsis* sp. using a constraint-based approach. The aim is to reconstruct the metabolic network of the alga and to create a genome scale model from the reconstruction.

The thesis is structured as follow: a background/introduction to the research field in Section 1. This is followed by Section 2 where the Biological and System biological concepts are explained. Section 3 is a thorough explanation of methodologies and Section 4 presents the results and discussion. Section 5 described the conclusion and future work. Appendices and references are given in Section 6 and 7, respectively.

2 Theory and literature review

2.1 Biological principles

2.1.1 Metabolism

Metabolism is the collection of all enzymatic reactions that are required for the growth, development and division of cells [24] . Metabolism is divided into two processes: anabolism and catabolism (Figure 1) [24] [25]. **Anabolism** is the process where organisms take up nutrients from the environment, and uses the nutrients as building blocks in order to build larger and more complex molecules [24] [25] . Anabolism requires energy, and leads to the production of new chemical compounds that subsequently function as precursors for cell components such as the synthesis of lipid, an essential cellular component in alga. Anabolism is therefore essential for cellular growth. The processes involved in anabolism could also be called biosynthesis processes as a collective term. Biosynthesis of all important molecules in the cells requires energy. So in order to produce new cell components, cells also need to be able to produce energy. In similar processes to uptake of nutrients, energy sources are also taken up from the environment. The two sources that provide energy for a cell are light and energy rich chemical compounds [24]. Organisms like plants and algae use light to generate ATP and NADPH, which are molecules with high chemical energy, through photosynthesis. Plants and algae are hence called phototrophic organism. Humans and animals, on the other hand, use energy being produced from degradation of chemical compounds such as sugar. Organisms generating energy this way are referred to as chemotrophic. These processes, breaking down larger molecules in order to generate energy, are called **catabolism**.

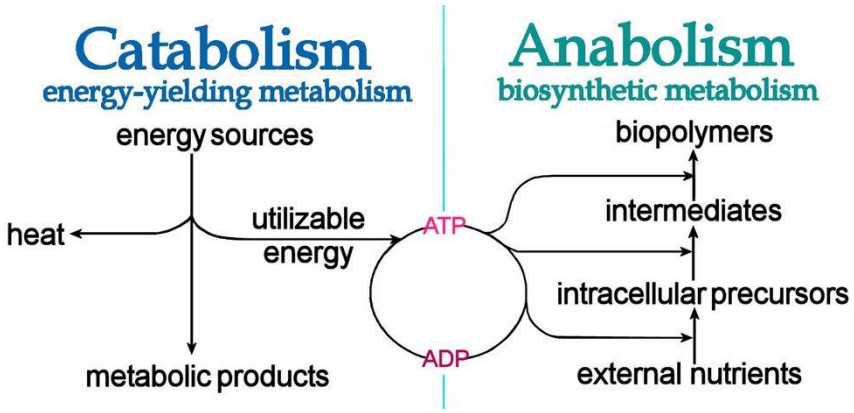


Figure 1. The two process involved in metabolism. Metabolism can be divided into two processes: catabolism and anabolism. Catabolism produces energy by breaking down nutrients in order to produce energy, while anabolism consumes energy to synthesize precursors that make up the cell. Figure retrieved from http://krebbing.blogspot.no/2007/12/anabolism-and-catabolism.html

Four main elements essential for the production of chemical compositions of the cell are carbon, oxygen, hydrogen and nitrogen. They are the main building blocks for all macromolecules such as DNA and RNA, proteins, lipids and polysaccharides. Other elements may also be necessary, but in smaller amounts. They are referred to as trace elements due to their low concentration in the cell. Common trace elements are phosphate, potassium, calcium, magnesium, sulfur, iron, zinc, manganese, copper, molybdenum, cobalt. Even though they are present in a smaller amount, they are still very important for a normal cell function. All these elements have to be supplied through the environment.

Metabolic reactions generally require specific enzymes to catalyze them [24]. Enzymes are proteins with a catalytic function, coded by the organism's genome. Usually, each enzyme catalyzes one specific reaction and only accepts one specific substrate molecule. Enzymes are not consumed during reactions, and they are available for further reactions. In database such as BRENDA [26], METACYC [27] or KEGG [28], the enzymes are identified as EC numbers, a code specifically referring to one enzyme or a group of enzymes having a particular function.

2.1.2 Metabolic pathways

A metabolic pathway is a sub-network of an entire metabolism that connects a group of metabolites to other metabolites through a group of reactions [24] [25]. Each metabolic pathway contains a group of biochemical reactions that are linked by their intermediates. It means in a

metabolic pathway, upstream reactions produce substrates for downstream reactions. The flow in metabolic pathways is often considered as unidirectional. Metabolic pathways play particular roles in the cell, for example, produce energy in glycolysis (catabolism) or generate monomers precursors in anabolism (purine biosynthesis). There are several pathways in the metabolism, such as purine biosynthesis, glucoronate metabolism, pentose phosphate pathway, to name a few. Among them, several metabolic pathways are referred as the core pathways. They are those pathways involving the core nutrients such as carbohydrate, fatty acid and amino acids [29]. Those pathways, such as the glycolysis, the tricarboxylic acid (TCA) cycle, the pentose phosphate pathway and the Calvin cycle are referred as core pathways. It is generally accepted that the core metabolic pathways are almost the same in all organisms [24].

2.1.3 Energy generation

Anabolic metabolism and transport reactions require energy. This energy is generated through catabolism, when nutrients are broken down to generate ATP and NADH. However, the energy production process produces heat, and if it was done in a single step, too much heat will be produced and burn the cell [24]. Hence, a nucleotide adenosine triphosphate (ATP) is formed as a carrier molecule in a stepwise energy generation process.

ATP can either be produced by photosynthesis, oxidative phosphorylation carried out by mitochondria, or by substrate-level phosphorylation [24]. In the first two processes, an enzyme called ATP synthase generates ATP by utilizing a proton gradient across a membrane. Whereas in the latter, ATP is produced by certain reactions involved in the core metabolism, where ADP acts as a cofactor. In these reactions; ADP directly receives a phosphate group and thus becomes ATP. During the conversion of phosphoenolpyruvate to pyruvate, for instance; an enzyme called pyruvate kinase transfers a phosphate molecule from phosphoenolpyruvate to ADP, resulting in pyruvate and ATP:

$$\text{Posphoenolpyruvate} + \text{ADP} + H^{+} \rightarrow \text{Pyruvate} + \text{ATP}$$

This reaction also occurs in fermentation process in some organism. The overall energy these organisms are able to accumulate through ATP production is smaller than the one of organisms performing respiration or photosynthesis.

2.1.4 Nannochloropsis

Nannochloropsis is a unicellular marine phytoplankton [30]. It belongs to the class of Eustigmatophyceae, consisting of various brown algae and diatoms. The group of Eustigmatophyceae containing *Nannochloropsis* is the Heterokontophyta [31]. This alga is assumed to have been created through a secondary endosymbiotic event, which is the most likely explanation for its plastid being surrounded by four membranes [32]. Six species of *Nannochloropsis* are recognized, including *N. gaditana*, *N. granulata*, *N. limnetica*, *N. oceanica*, *N. oculata* and *N. salina*. Among them, genome sequences of *N. gaditana* and *N. oceanica* are available.

Comparision between the two genomes show that out of 6,395 orthologous groups (OGs) identified, 5,048 OGs are present in both *N. oceanica* and *N. gaditana* [33]. OGs are genes that has evolved from the same ancestor. The high number of OGs found in both species indicates the two genomes have been conserved during the course of evolution. The presence of species-specific genes may reveal a difference in biological functions between *N. gaditana* and *N. oceanica* [33]. However, the finding of A. Vieler and coworkers as well as R. Radakovits and coworkers have highlighted the similarity of the central metabolic functions for these two species [33] [21]. Their core metabolic pathways are predicted to be the same. The main differences is the number of putative enzymes. For instance, *N. gaditana* contains six putative carbonic anhydrases, which is an essential enzyme that catalyzes the reversible hydration of carbon dioxide, while only two putative enzymes has been found in *N. oceanica* [21]. This inconsistency might reflect that *N. gaditana* is better adapted to a low availability of inorganic carbon sources.

This metabolic model has been constructed and curated based on the metabolic information of both of these two *Nannochloropsis* species. As the genome sequence of *N. gaditana* is more complete and has a higher degree of annotation than that of *N. oceanica* the draft reconstruction was generated from the *N. gaditana* genome. The estimated genome size is about 29 Mb which covers a predicted number of 9052 genes.

Nannochloropsis is an alga with several possible commercial applications due to its relatively rapid growth rate, and its ability to produces and store high amounts of lipids and also several commercially valuable pigments, both which can be converted into other useful bioproducts. *Nannochloropsis* is widely used as feedstock for aquaculture hatcheries due to its valuable

nutrient contents such as polyunsaturated fatty acids [34]. In addition, *Nannochloropsis* is able to accumulate up to 60 % lipids (compared to its own weight) under conditions favourable for lipid synthesis [33] [17]. Compared to other lipid producers such as palm tree and *Chlorella* sp., the lipid production of *Nannochloropsis* is much higher (27000 kg/ha/year) while the other two organisms yield 5000 kg/ha/year and 10,000 kg/ha/year, respectively [33].

2.2 Systems biology principles

Systems biology has emerged as a novel approach of studying metabolic pathways, regulatory networks and signal transduction networks [11]. Out of the many computational approaches that have been developed within these fields (overview given in Figure 2), constraint-based modeling is the most frequently applied technique [11].

Method	Model systems	Parameterization	Typical prediction type	Advantages	Disadvantages	Refs
Stochastic kinetic modelling	Small-scale biological processes	Detailed kinetic parameters	Reaction fluxes, component concentrations and regulatory states	• Mechanistic • Dynamic • Captures biological stochasticity and biophysics	• Computationally intensive • Difficult to parameterize • Challenging to model multiple timescales	106
Deterministic kinetic modelling	Small-scale biological processes	Detailed kinetic parameters	Reaction fluxes, component concentrations and regulatory states	• Mechanistic • Dynamic	• Computationally intensive • Difficult to parameterize	107
Constraint-based modelling	Genome-scale metabolism	Network topology, and uptake and secretion rates	Metabolic flux states and gene essentiality	• Mechanistic • Large scale • No kinetic information is required	• No inherent dynamic or regulatory predictions • No explicit representation of metabolic concentrations	3,104
Logical, Boolean or rule-based formalisms	Signalling networks and transcriptional regulatory networks	Rule-based interaction network	Global activity states and on–off states of genes	Can model dynamics and regulation	Biological systems are rarely discrete	108
Bayesian approaches	Gene regulatory networks and signalling networks	High-throughput data sets	Probability distribution score	• Non-biased • Can include disparate and even non-biological data • Takes previous associations into account	• Statistical • Issues of over-fitting • Requires comprehensive training data	109, 110
Graph and interaction networks	Protein–protein and genetic interaction networks	Interaction network that is based on biological data	Enriched clusters of genes and proteins	• Incorporates prior biological data • Encompasses most cellular processes	• Dynamics are not explicitly represented	111, 112
Pathway enrichment analysis	Metabolic and signalling networks	Pathway databases (for example, KEGG, Gene Ontology and BioCyc)	Enriched pathways	• Simple and quick • Takes prior knowledge into account	• Biased to human-defined pathways • Non-modelling approach	73

KEGG, Kyoto Encyclopedia of Genes and Genomes.

Figure 2. Different modelling and analysis techniques for high throughput data Constraint based modelling is the most common used approach to study the genome scale metabolic model since it does not require kinetic data which is still limited at the moment. The Figure was obtained from Bordbar, A., et al. [11]

2.2.1 Constraint-based models (CBMs)

Organisms' development depends on nutrients and growth conditions. And because the fluxes of nutrients into a cell are limited, these fluxes provide an upper limit to all reaction fluxes in the model, as all the compounds and reaction in a model are usually linked together. Constraint based methods utilize these cellular flux limitations of the biological organisms to study their behaviors [11] [35].

In general, a metabolic network describes which products are produced from a particular substrate [35]. For example, Figure 3 shows the simplified metabolic network of ethanol production from sugar [35]. The boxes A, B and C represent pools of the compounds involved in the production process. The orange arrows describe the fluxes of the exchange reactions, transporting substrate or product in or out of the cell via the cell membrane. The green arrows show the internal fluxes.

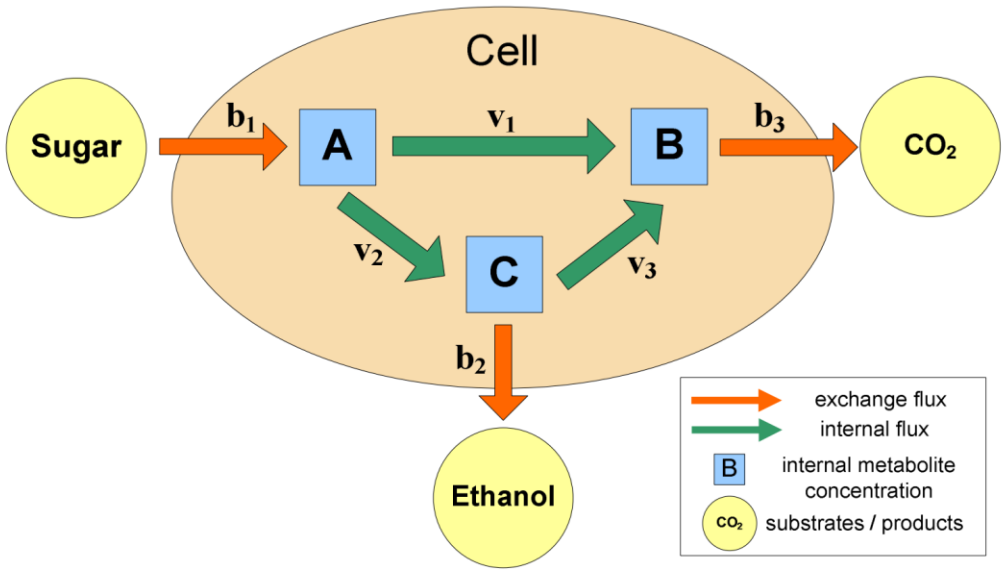


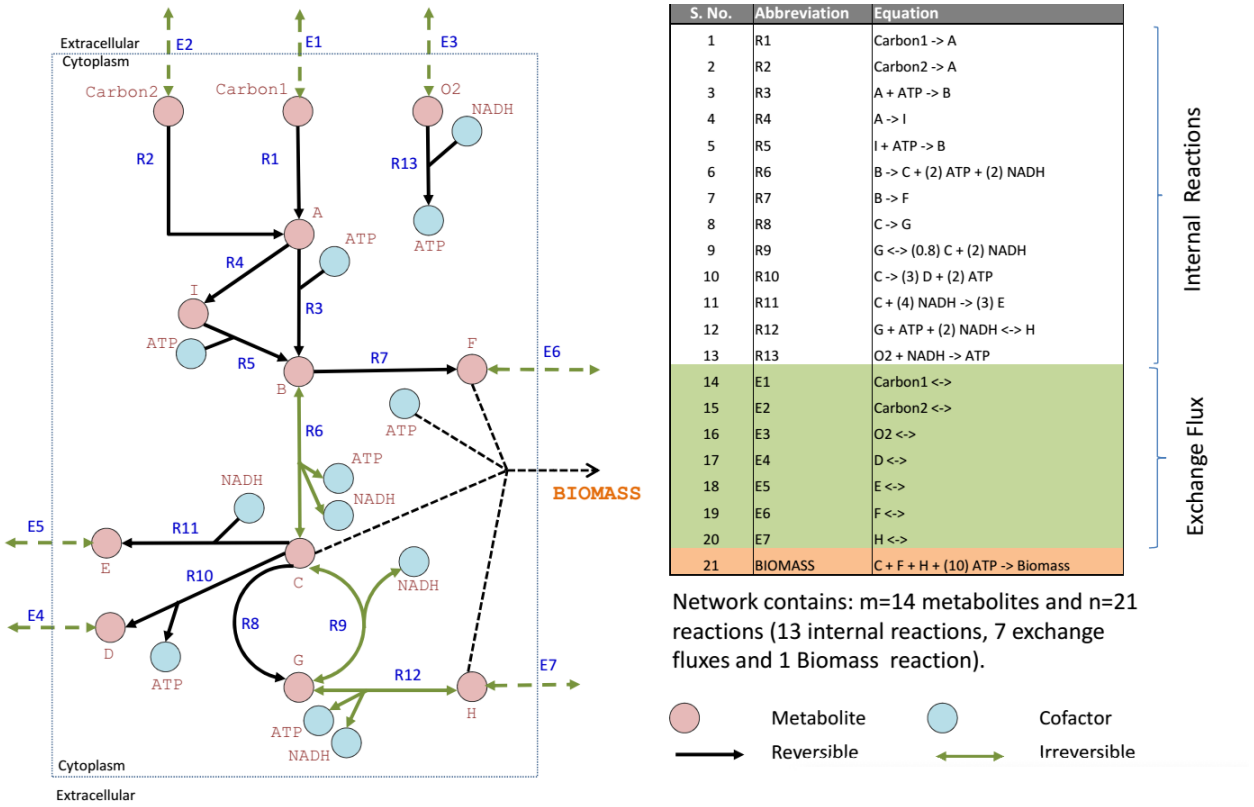
Figure 3. Simplified metabolic network to produce ethanol from sugar. The cell takes up sugar and produces carbon dioxide and ethanol via several reactions. The Figure was obtained from Samal, A. [35]

Ideally, when describing what is going on inside a cell, one would need to quantify each reaction flux in the cell. It is relatively easy to determine fluxes of exchange reactions [35]. These usually include the transport of sugar, oxygen or other nutrients that are required for the core metabolism of the organism, and the export of other products, such as carbon dioxide and ethanol [35]. The most difficult task is to determine the internal reaction fluxes [35].

The kinetics, and also the concentration of enzymes and cofactors greatly affect the enzyme activity, and this information is required when utilizing most of the current metabolic modelling methods. However, regardless of the rapid increase in availability of biological information, there is still a lack of mathematical data making it possible to model all the cellular metabolic processes taking place in a single cell. To work around this obstacle, several methods have been developed for different analysis purposes (Figure 2). Among these methods, the most useful approach utilizing the constraints of a metabolic network to predict fluxes of a metabolic network is flux–balance analysis [35].

Flux based analysis has emerged as an effective approach to be used in metabolic modelling, because it only requires stoichiometric information [9]. Despite of the lack of detailed kinetic information, the study of fluxes through these networks is essential. The approach provides a quantitative study of the net flux distribution in the pathway that contributes to the effect of overall cellular functions. Therefore, analyzing the metabolic fluxes is a usual way to study the metabolic genotype-phenotype relationship. In fact, effective applications of constraint-based analysis of metabolic models have been proven in many fields, such as metabolic engineering, prediction of gene deletions response, and drug target identification [9].

Figure 4 shows a simple model that imitates the core metabolism of a metabolic network. The metabolic model can be converted into a mathematical model where the main component is a stoichiometric matrix (normally abbreviated an S-matrix) (Figure 5) [36] [35]. In such a stoichiometric matrix, the number of rows corresponds to all the metabolites that are involved in the system, while the number of columns corresponds to reactions taking place. A negative coefficient in the S-matrix represents a metabolite that is consumed in a certain reaction. In contrast, positive stoichiometric coefficients represent metabolites that are being produced. The dimension of the matrix is m x n where m is the number of metabolites, and n is the number of reactions. For example, the matrix in Figure 5 is a 14 x 21 matrix, with fourteen metabolites and twenty-one reactions.



S. No.	Abbreviation	Equation
1	R1	Carbon1 -> A
2	R2	Carbon2 -> A
3	R3	A + ATP -> B
4	R4	A -> I
5	R5	I + ATP -> B
6	R6	B -> C + (2) ATP + (2) NADH
7	R7	B -> F
8	R8	C -> G
9	R9	G <-> (0.8) C + (2) NADH
10	R10	C -> (3) D + (2) ATP
11	R11	C + (4) NADH -> (3) E
12	R12	G + ATP + (2) NADH <-> H
13	R13	O2 + NADH -> ATP
14	E1	Carbon1 <->
15	E2	Carbon2 <->
16	E3	O2 <->
17	E4	D <->
18	E5	E <->
19	E6	F <->
20	E7	H <->
21	BIOMASS	C + F + H + (10) ATP -> Biomass

Internal Reactions

Exchange Flux

Network contains: m=14 metabolites and n=21 reactions (13 internal reactions, 7 exchange fluxes and 1 Biomass reaction).

Metabolite

Cofactor

Reversible

Irreversible

Figure 4. Example of a core metabolic network depicting a system consisting of fourteen metabolites and twenty-one reactions. On the right, a core metabolic network that produces biomass was displayed. Figure on the left were all reactions that taken from the core metabolic network. The Figure was obtained from Samal, A. [35]

Stoichiometric Matrix S

	Internal Reactions												Exchange Flux							
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	E1	E2	E3	E4	E5	E6	E7	Biomass
Carbon1	-1	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0
Carbon2	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0
A	1	1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	1	0	1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	1	0	-1	0.8	-1	-1	0	0	0	0	0	0	0	0	0	-1
D	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	-1	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	-1	0	0	0
F	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	-1	0	-1
G	0	0	0	0	0	0	0	1	-1	0	0	-1	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	-1	-1
I	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O2	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	-1	0	0	0	0	0
ATP	0	0	-1	0	-1	2	0	0	0	2	0	-2	1	0	0	0	0	0	0	0	-10
NADH	0	0	0	0	0	2	0	0	2	0	-4	-1	-1	0	0	0	0	0	0	0	0

Figure 5. Example of a stoichiometric matrix. This particular stoichiometric matrix corresponds to the network depicted in Figure 4. The Figure was obtained from Samal, A. [35].

The rate of change in the concentration of a certain metabolite over time can be described as

$$\frac{dX}{dt} = S \cdot v$$

Here, X is the vector of metabolite concentrations, having a length equal to the number of metabolites in the system (m); S is the stoichiometric matrix of dimension m x n, and v is the flux vector, having a length equal to the number of reactions (n) (Figure 6).

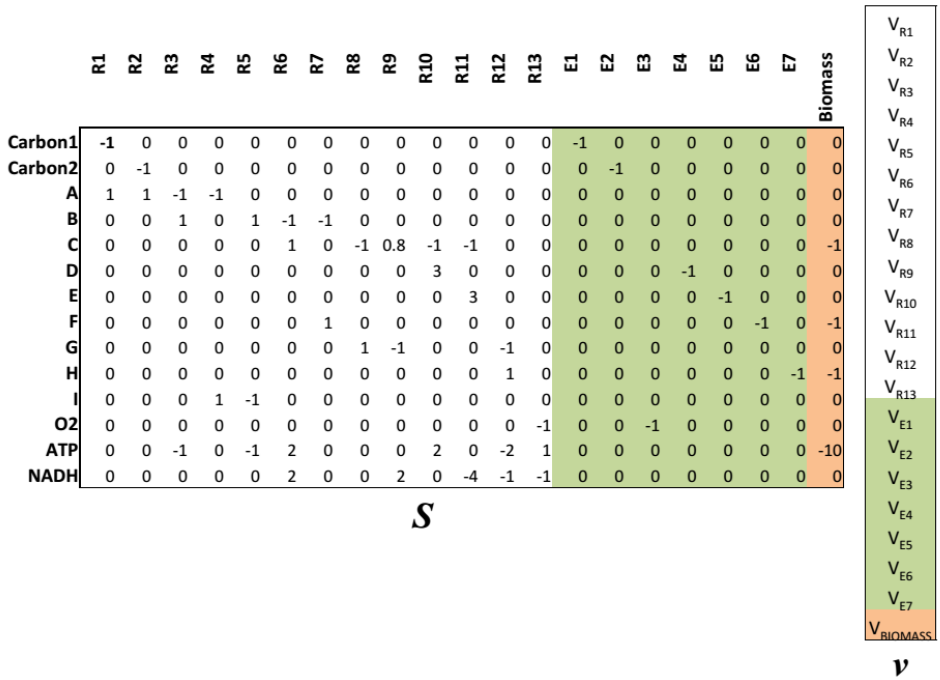


Figure 6. A system to describe the rate at which the metabolite concentration is changed. Here, S is a stoichiometric matrix, and v is the flux vector, where every flux has the denomination change in metabolite concentration per time unit. Each value in v represents the flux through a certain reaction. The Figure was obtained from Samal, A. [35].

At steady state, the rates of change in the metabolites concentration are zero for all reactions, meaning that all metabolites are consumed with the same rate as they are produced. Mathematically, this can be written as:

$$\frac{dX}{dt} = S \cdot v = 0$$

This is called the stoichiometric mass-balance constraint.

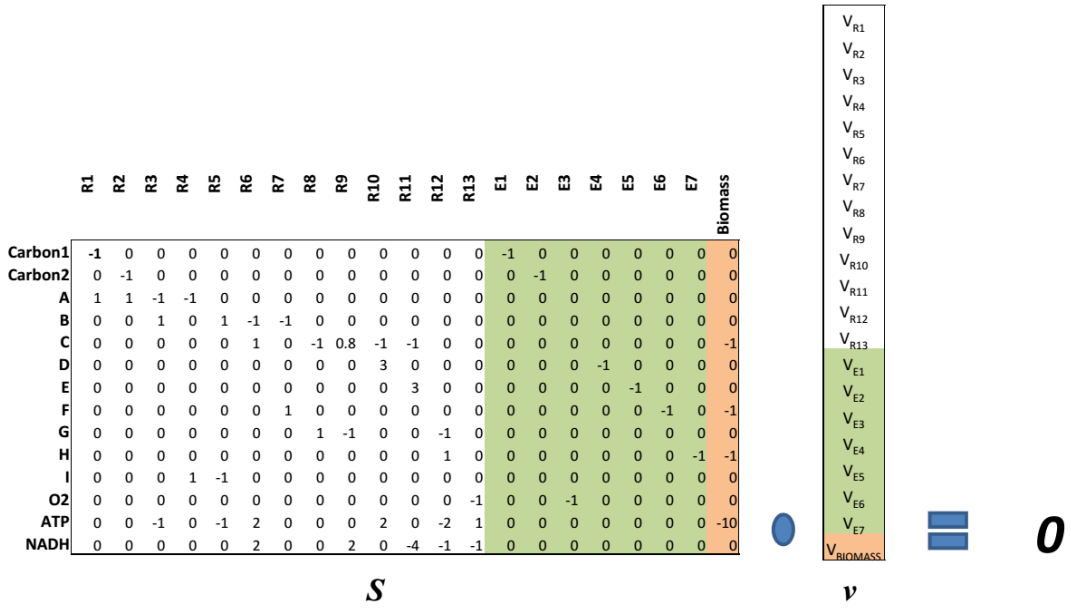


Figure 7. The stoichiometric mass balance of the network depicted in Figure 4 at steady state. The rate of which the metabolites are being consumed is then equal to the production rate. The Figure was obtained from Samal, A. [35].

The stoichiometric mass-balance yields a system of linear equations, describing every reaction flux of the system. The concentration of each metabolite will then at steady state be described by a linear equation, instead of a differential equation, which is the case before steady state is reached. For instance, for the metabolite A in the third row in the Figure 7, the linear equation describing the metabolite's concentration at steady state would be

$$v_{R1} + v_{R2} - v_{R3} - v_{R4} = 0$$

The dot product of the matrix and the flux vector in Figure 7, will then yield a system of linear equation as follows (Figure 8)

$$-V_{R1} - V_{E1} = 0$$
$$-V_{R2} - V_{E2} = 0$$
$$V_{R1} + V_{R2} - V_{R3} - V_{R4} = 0$$
$$V_{R3} + V_{R5} - V_{R6} - V_{R7} = 0$$
$$V_{R6} - V_{R8} + 0.8\ V_{R9} - V_{R10} - V_{R11} - V_{BIOMASS} = 0$$
$$3\ V_{R10} - V_{E4} = 0$$
$$3\ V_{R11} - V_{E5} = 0$$
$$V_{R7} - V_{E6} - V_{BIOMASS} = 0$$
$$V_{R8} - V_{R9} - V_{R12} = 0$$
$$V_{R13} - V_{E7} - V_{BIOMASS} = 0$$
$$V_{R4} - V_{R5} = 0$$
$$-V_{R13} - V_{E3} = 0$$
$$-V_{R3} - V_{R5} + 2\ V_{R6} + 2\ V_{R10} - 2\ V_{R12} + V_{R13} - 10\ V_{BIOMASS} = 0$$
$$2\ V_{R6} + 2\ V_{R9} - 4\ V_{R11} - V_{R12} - V_{R13} = 0$$

Figure 8. Linear systems of the whole S matrix describing metabolite concentrations by using reaction fluxes at steady state by Samal, A. [35].

The system described in Figure 8 contains 14 equations for 21 fluxes. When compared to the m x n stoichiometric matrix in Figure 8, it means that the number of equations (which is the same as the number of metabolites, m) is less than the number of unknown fluxes (which is the same as the number of reactions, n). This is a necessary prerequisite for finding a solution to the fluxes.

The linear system above (Figure 8) gives a vastness of possible results. Any flux vector ν that satisfies the system above is said to be in the null space of S, hence is a possible solution [9] [35].

In reality, metabolism is limited by several constraints which in practice will decrease the size of the solution space [9]. There are three kinds of constraints that need to be addressed. The first one is thermodynamic constraints, which limit certain reactions to being irreversible [35] [12]. For an irreversible reaction j, its flux is bounded in one direction, so $\nu_j \geq 0$. The second constraint is enzyme capacity, which restrict the upper bounds of certain reactions [35] [12]. So for a reaction j, the enzyme capacity restricts the flux to $\nu_j \leq \nu_{max}$. The last constraint is environmental limitations, which determines the lower bound of certain reactions due to the limited of nutrients in the environment [35] [12]. For instance, if a specific nutrient is not in the environment, then the lower bound of the reactions that use this specific nutrient is set to zero. The upper bound is not constrained in terms of secretion [35]. These three constraints limit the size of the solution space as can be seen in Figure 9.

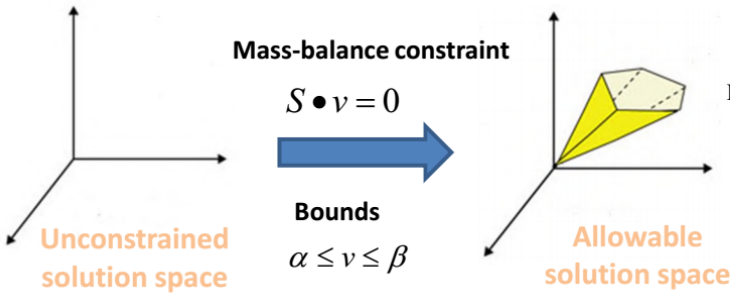


Figure 9. The addition of constraints limit the solution space. Constraint such as enzyme capacity restricts the flux of the reaction into a certain value $\alpha \le v \le \beta$. The Figure was obtained from Samal, A. [35].

In general, the purpose of metabolic modeling is to study specific phenotype responses under different external conditions. In such application, a cellular objective is set to determine a particular flux distribution in the solution space. The most common objective function is to maximize the biomass. This is also the situation that best describes reality, as cells are mainly "interested" in multiplying, in order to create new cells. In an actual cell, new biomass is not created in one reaction, but in a multitude of different reactions, all producing molecules that are needed for cell growth. But in order to describe cell growth in a metabolic model, an artificial biomass reaction which consumes the molecules necessary for "building" new cells in an experimentally measured stoichiometric ratio, is added to the model and set as objective function [37].

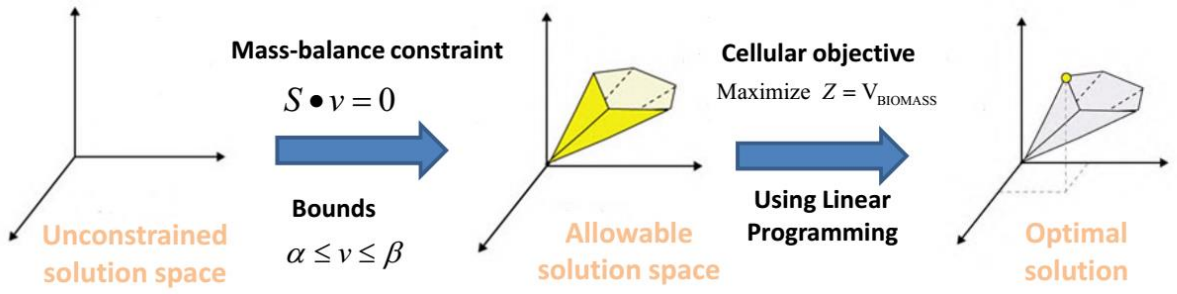


Figure 10. Determination of a particular flux distribution in a suitable solution space. Linear program was used together with a cellular objective. More often the objective is to maximize biomass. In this case, an artificial biomass reaction which consumes the molecules necessary for "building" new cells in an experimentally measured stoichiometric ratio is added. The flux through this reaction is equal to the exponential growth rate of the organism. The Figure was obtained from Samal, A. **[35].**

Figure 10 illustrates a linear problem meant to determine the highest allowed flux through the biomass equation, limited by the constraints of the system. The problem can be described as

Solve S . v = 0 where $\alpha \leq v \leq \beta$

To obtain v that : maximizeZ= vbiomass

In this case, flux balance analysis can be utilized to solve the system of unknown fluxes, and the optimal solution is given as Figure 11. As shown in the Figure, the red flux indicates the optimal flux in order to obtain the maximal flux through the biomass reaction.

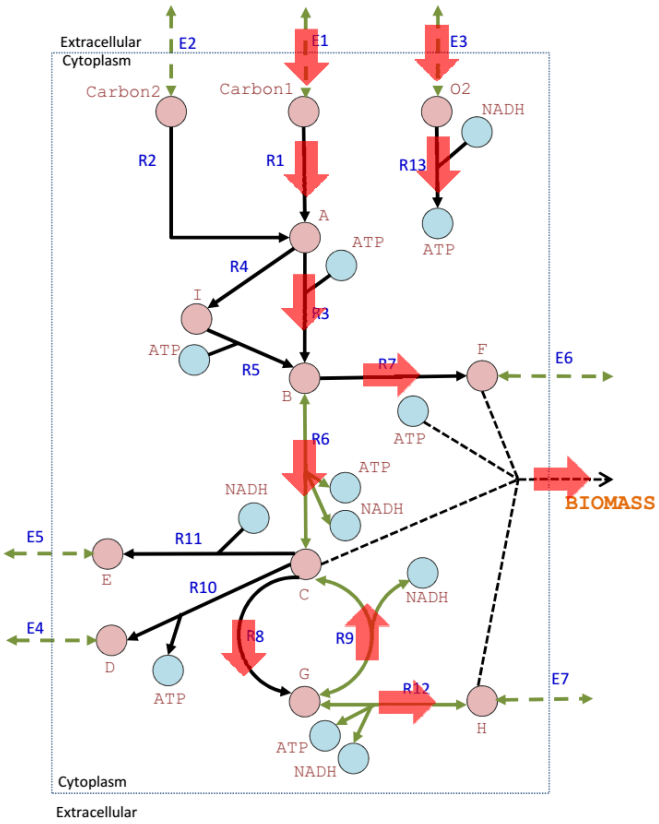


Figure 11. Optimize fluxes (red) to obtain maximize the biomass production. The diagram shows the optimize flux that will produce maximize biomass concentration. The Figure was obtained from Samal, A. [35].

Prediction of fluxes from well curated constraint-based metabolic models has proved to share consistency with experimental data gathered for the same system [38] (Figure12).

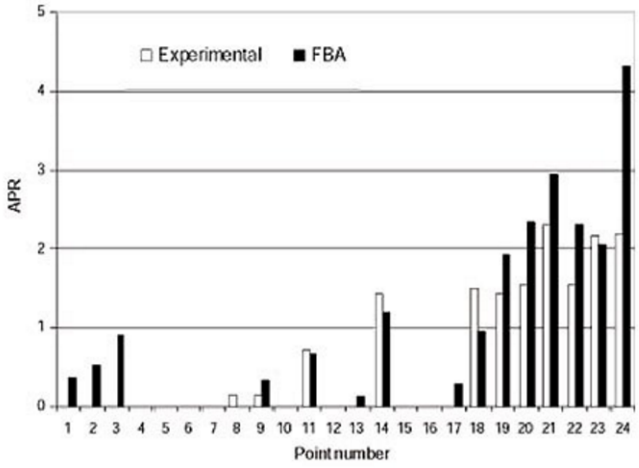


Figure 12. Optimal fluxes predicted by flux balance analysis performed on an *in silico* well curated *E. coli* model. The flux balance analysis (FBA) data show that the simulation results are consistent with the experimental data. The Figure was obtained from Edwards, J.S., R.U. Ibarra, and B.O. Palsson [38].

However, no kinetic parameters such as enzyme concentrations and enzyme conversion rates are used in flux balance analysis. This approach only considers the metabolism at steady state, so it cannot foresee the internal fluxes [35]. In addition, the simplest form of flux balance analysis does also not consider gene regulation, so its predictions are not always correct [35]. But most of the time, they offer a good prediction of the fluxes in the system.

2.2.2 Recent development of constraint- based models (CBMs)

Constraint-based analysis has been used in metabolic reactions networks for over 27 years [11]. In the early years of development, CBMs were mainly used to analyze the pathway productivity and metabolites overflow. Study of Fell, D.A. and Small J.R. in 1986 [39] were among the very first publication used CMBs to study the fatty acid synthesis from glucose. In 1989, R. A. Majewski and M.M.Domach [40] were the first scientists study the production of acetate in E.coli by CBMs. The next decades witnessed the significant development of whole genome sequencing techniques, which facilitated construction of CBMs at the genome-scale, and allowed for simulation of organisms' entire metabolism in addition to phenotypes. Since the reactions in these early CBMs had direct links to the corresponding genes coding the proteins catalyzing the

reactions, studies of the effect of gene knock-out was the most frequent application of CBMs [11]. In addition, these CBMs also enable study about cellular behaviors in a comprehensive manner such as pathway structures, metabolic fluxes and bacterial evolution. As a result of the growth of cheaper and more reliable high through-put techniques in the following years (2005-2009), CBMs were built using 'omics' data, in order to increase the context specificity of these models [11]. In 2007 Oh, Y.-K., and coworkers [41] presented a genome scale (GNS) reconstruction of *Bacillus subtilis* metabolism. The model was built based on the combination of genomic, biochemical, and physiological information and high-throughput phenotyping experiments. The final model covered 844 ORFs, 1020 reactions and 988 metabolites. In 2009, a new GNS model of *Bacillus subtilis* 168 was presented [42]. The model was developed based on the most accurate *B. subtilis* 168 genome annotation at that time by the SEED. The reconstruction was among the most complete model of *B.subtilis* available with 1,103 genes and 1,437 reactions. In 2009, Boyle, N.R. and Morgan J.A. [12] introduced the first model for algae, the flux balance analysis model in *Chlamydomonas reinhardtii* which include three metabolically active compartments with central and intermediary metabolism. The reconstructed model contains 484 reactions and 458 intracellular metabolites. The model was used to analyze metabolic fluxes under autotrophic, heterotrophic and mixotrophic growth using FBA.

From 2010 until now, big efforts have been put into generating CBMs of higher quality, in order to provide more meaningful biological interpretations. Genome-scale models have also been reconstructed for many species from all the three main kingdoms Archea, Bacteria and Eukaryota. The majority of metabolic reconstructions have been focusing on bacteria. However, the number of reconstructions that have been performed for algae and plants are still limited [11] [8] [14]. There were only *Arabidopsis thaliana* for plant and *Chlamydomonas reinhardtii* for algae have been modeled at the genome scale level.

Another study of primary metabolism (which includes core metabolic pathways) was performed by de Oliveira Dal'Molin, C.G. and coworkers in 2010 [14] . They built a genome-scale model of the core metabolism of *Arabidopsis*. The model was created from the *Arabidopsis* genome annotation and included 1,419 unique open reading frames, 1,748 metabolites, 5,253 gene-enzyme reaction-associations and 1,567 reactions, both located in the cytoplasm, the mitochondria, the plastid, the peroxisome, and the vacuole. This model describes the pathway of photo respiratory in addition to important differences between redox metabolism in

photosynthetic and non-photosynthetic plant cells. The reconstruction can serve as a framework making it easier to study functional analysis as well as being used for generating novel knowledge about the metabolism of plants.

In 2014, Elena Vinay-Lara and coworkers [43] constructed a GNS model to compare metabolic difference between two strains of *Lactobacillus casei* (ATCC 334 and 12A). The draft models were generated from the RAST genome annotations system utilizing the ModelSEED database. The model was then used to evaluate ATP production, mass-and-charge-balance of reactions, and growth phenotypes of the two *L. casei* sub-species. The authors found that the metabolism was the same in two strains of *L. casei*. The *L. casei* ATCC 334 model contains 1,040 reactions, 959 metabolites and 548 genes, while the *L. casei* 12A model covers 1,076 reactions, 979 metabolites and 640 genes. Also, in 2014, a GNS model describing *Synechococcus elongates* was developed by Julián Triana and coworkers [7]. The reconstruction includes 851 reactions and 838 metabolites.

Much research also focus on refining already existing model in the metabolic model databases such as BiGG, by incorporating new compartments, reactions and transporters. For example, the existing GNS metabolic model of rice leaf was curated and studied by Chatterjee, A. and Kundu S. in 2015 [44]. They used the re-curated model to explore chlorophyll synthesis in a leaf. Crucial reactions involved in chlorophyll synthesis, and their associated genes, were predicted and validated against the existing experimental data.

In the recent year the study direction of GNS was changed to either curate the existing genome-scale metabolic models and applied them to explore biological capacities or to reconstruct model for less-characterized organisms. For instance, Hendry, J.I., [45] generated an updated GENRE for existing *Synechococcus sp*. PCC 7002 model in 2016. They performed flux balance analysis and double gene knock outs to get insights into the possibilities of performing genetic engineering on this organism. Also in the beginning of this year 2016, Levering, J. and coworkers [46] built a novel model for *Streptococcus pyogenes* M49 with 480 genes associated with 576 reactions and 558 metabolites. Almost at the same time, Vongsangnak, W. and colleagues [47] created a genome-scale metabolic model of an oleaginous fungus, *Mucor circinelloides* iWV1213, considered for industrial use. Their model contained 1213 genes, 1413 metabolites and 1326 reactions taking place in various compartments.

2.2.3 Construction of constraint-based models

The process of network reconstruction and modelling is time-consuming, and could take in total from some months to several years. When modeling a network, it is a good approach to combine the semi-automatic tools with manual refinement. An organism having a genome with a poor degree of annotation would require more manual curation. Therefore, the time frame for making a reconstruction of an organism's metabolism could vary, depending on the accuracy of the metabolic data available for the organism in question.

Several researchers have focused on tools for generating genome-scale metabolic reconstructions [8] [45] [37] [41] [1]. In general, all the developed methods or tools perform four basic steps, described in Figure 13. First, a draft reconstruction is produced semi-automatically. Secondly, a manual reconstruction refinement is required. Thirdly, the model is converted to a mathematical representation. And finally, researchers need to debug and validate the model based on experimental data collected for the organism being modeled.

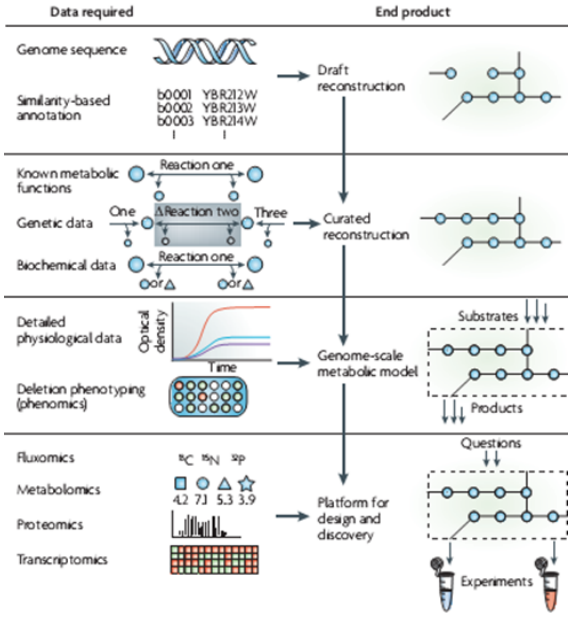
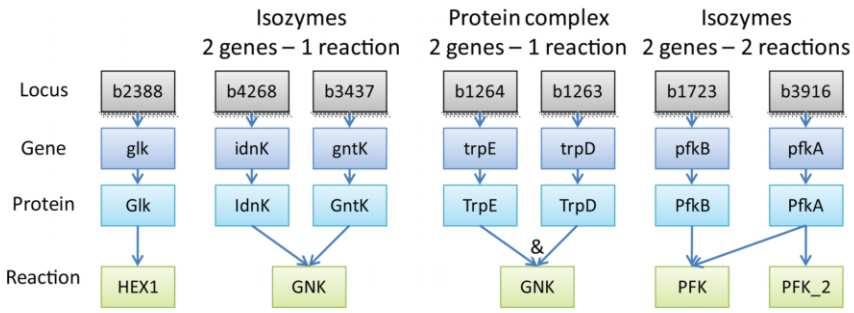


Figure 13. The four steps required to generate a metabolic model. First, a draft reconstruction is generated from the target organism's genome. Then, the reconstruction is curated based on knowledge about organism's metabolism or genomic/biochemical data. In the next step, the reconstruction is converted into a mathematical model. Debugging and validating the model are also performed within this step. Finally, the finalized model is ready to serve as a platform for design and discovery. The Figure was obtained from Feist, A.M., et al [37].

2.2.3.1 Creating a draft reconstruction

Creating draft reconstructions relies on having access to an adequate annotation of the genome of the organism to be modeled. Annotated genomes can often be found in online databases. Genome annotations can be retrieved in organism-specific databases, such as EcoCyc [48] for *Ecoli*, SGD (*Saccharomyces* Genome Database) [49] , or CYGD (Comprehensive Yeast Genome Database) [50] for *Saccharomyces cerevisiae*. General databases such as EntrezGene [51], CMR (comprehensive Microbial Resource) [52], Genome Reviews (through EBI; European Bioinformatics institute) [53] or IMG (Intergrated Microbial Genome) [54] could also be used. The unique characteristic of an organism metabolic reconstruction is determined by the genome annotation from its genome. As it can provide specie-specific information about the metabolic enzymes that could be in the target organism [37].

Metabolic functions are then obtained by combining information about genotype and phenotype. This "marriage" results in a list of Genes-Proteins-Reactions Associations (GPRs) which represents the reactions catalyzed by the enzymes present in the genome of the organism in question (Figure 14). Metabolic functions were extracted from metabolic databases, biochemical data or literature. Databases such as KEGG (Kyoto Encylopedia of Genes and Genomes), BRENDA (Braunschweig Enzyme Database) and MetaCyc contain information about metabolic reactions that has been proven to present in numerous organisms. In most enzyme driven reactions having been associated with one or more organisms, the enzyme has been linked to an Enzyme Commission (EC) Number or a Transport Commission (TC) number. It is possible to extract information from these databases manually by checking every enzyme and reaction of a specific organism, or the process could be automated, by utilizing already developed tools to collect reactions for a certain organism from the metabolic databases.



Reaction Abbreviation	Reaction Name	E. C. number	GPR
HEX1	hexokinase (D-glucose:ATP)	2.7.1.1	(b2388)
GNK	gluconokinase	2.7.1.12	(b3437) or (b4268)
ANS	anthranilate synthase	4.1.3.27	(b1264) and (b1263)
PFK	phosphofructokinase	2.7.1.11	(b1723) or (b3916)
PFK_2	Phosphofructokinase (2)	2.7.1.11	b3916

Figure 14. Example of gene - protein-reaction associations for *E.coli*. Combining information about genotype and phenotype results in a list of Genes-Proteins-Reactions Associations (GPRs) which represents the reactions catalyzed by the enzymes present in the genome of the organism in question. The Figure was obtained from Thiele, I. and B.Ø. Palsson [1].

However, there may be a difference in substrate specificities and enzyme activities between enzymes with the same EC or TC number [37]. As the consequence, the actual enzymatic reactions in the target organism may distinct from that of the analogous enzyme in the reference organism. Additionally, there may be inadequate information about cellular compartment location and reaction directionality. This missing information together with other errors made during genome annotation, could make the draft reconstruction less accurate and cause it to contain gaps. Hence, draft models needs to be further refined before they are ready for simulation.

2.2.3.2 Model refinement

In this step, organism specific databases, textbooks, primary publications, review articles and experts opinions are the main data sources providing information about different aspects of reactions, such as compartment location and directionality. The subcellular compartment of specific metabolic reactions can be addressed using studies of protein localization. Likewise, information about reversibility and substrate specificity can be generated from biochemical studies. The purpose of this step is to fill the gaps that most likely exist in the reaction network, either by inference with biochemical data or through direct evidence from literature. The practical details of gap-filling are discussed in the material and method section.

Gaps could arise due to the establishment of incorrect gene-protein-reaction associations because of inconsistencies in the gene annotation process, but also from lacking information about enzymes (and thus reactions) present in the model [37]. The presence of gaps in metabolite pathway will create dead-end metabolites [55], which are metabolites that are either only produced or only consumed in the model. Since the concentrations of these metabolites can only increase or decrease, it is impossible to find a steady state solution for the fluxes in the reactions where these metabolites are involved. Since the only allowed steady-state solution for the fluxes of the reactions these metabolites participate in is 0, all these reactions will be block (Figure 15). Two classes of dead-end metabolites can be identified, namely Root-Non Produced metabolites, which are only consumed in the model, and Root-Non-Consumed metabolites which are only produced [55].

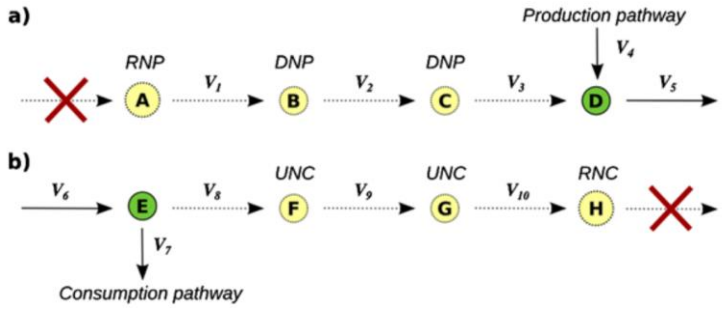


Figure 15. Classifications of gaps in a metabolic model. a) Root-non-produced (RNP A) gap is a metabolite that is only consumed and is not produced in any reaction in the model. This gap blocks the production pathway. b) Root-non-consumed (RNC H) gap is a metabolite that is only produced and is not consumed in any reaction in the model. The gap blocks the consumption pathway. The Figure was obtained from Ponce-de-León, M., Montero F., and Peretó J. [55].

Because of the need for gap filling, a high quality network reconstruction is not only based on automatic genome-based procedures, but also required a considerable time and effort spent performing detailed manual curation based on literature. In an attempt to decrease the time-consuming aspect of gap-filling, a biochemically, genomically, and genetically (BiGG) [56] structured knowledge based database that is organism specific and available to all researchers have been established.

2.2.3.3 Conversion from metabolic model to MATLAB model and debugging

In order to be able to perform network analysis by available computational tools, the genome-scaled network reconstruction needs to be converted to a mathematical model. The need for evaluating the network properties of constraint-based models in a mathematical form, has led to the development of COBRA, a COnstraint-Based Reconstruction and Analysis framework [57]. Genome scale metabolic reconstruction can be converted into a mathematical model by using COBRA toolbox combined with a platform such as MATLAB or Python [58]. MATLAB is a program language that has been used by a large number of scientists worldwide to analyze and design system. A mathematical metabolic model in MATLAB is a structure containing both vectors and matrices that together describes the network completely. An example of how a mathematical model in MATLAB could look is given in Figure 16 [1].

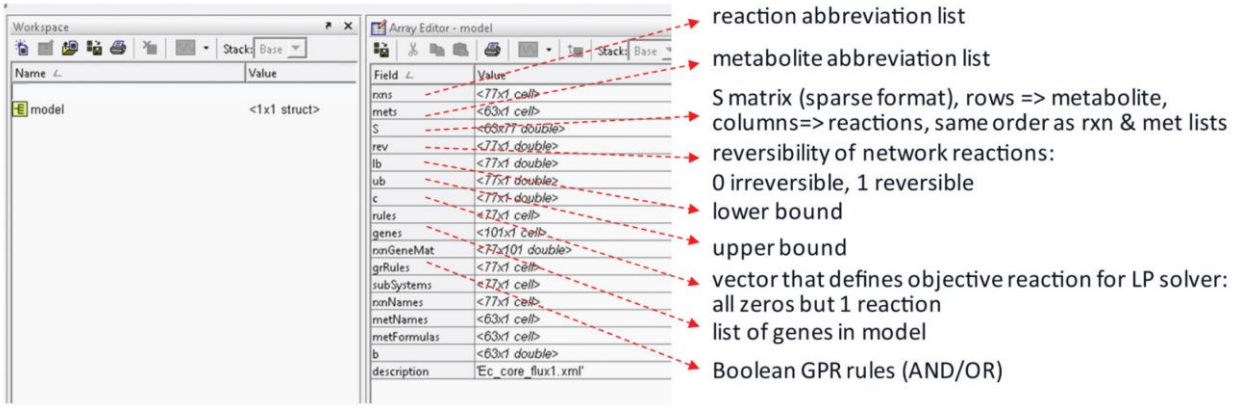


Figure 16. Main components of a mathematical metabolic model in MATLAB. A mathematical metabolic model in MATLAB is a structure containing both vectors and matrices that together describes the network completely. The Figure was obtained from Thiele, I. and B.Ø. Palsson [1].

In this stage, additional gap-filling could be performed if it is needed to fix incomplete pathways, or to remove reaction incorrectly included in the automatic draft reconstruction step [1].

2.2.3.4 Validation of the model

Mathematical models can be compared to phenotypic data in order to validate how accurately the model describe the modeled organism [1]. If the consistency between the phenotypes of the modeled organism and the mathematical model is poor, more manual curation may be necessary. Therefore, a repeated cycle of wet-lab and dry-lab research could be necessary to continuously improve the prediction of the model, as well as facilitating the generation of hypothesis [8].

3 Materials and method

The objective of this master project has been to build a genome-scale reconstruction of the metabolism in *Nannochloropsis* sp. using a constraint-based approach. This is done by creating a draft genome scale metabolic reconstruction using PlantSeed. Then the initial draft was converted to a mathematical model using MATLAB. Follow by the manual curation of the initial model to generate a genome scale model for the alga. In this section, tools and database as well as the method that used to create the genome scale model will be introduced.

3.1 Software environment

3.1.1 MATLAB [58]

MATLAB is a high-performance programming language, which covers computation, visualization and programming in a user-friendly interface. The platform is specialized to solve engineering and scientific problems. Among the most natural way to express computational mathematic is the matrix-based MATLAB language. In this project, MATLAB (version R2014b) was utilized as the main program in order to analyze the properties of the model, and also, to identify problems in the model, and parts that needed further refinement.

3.1.2 COBRA_toolbox [57]

COBRA toolbox is a MATLAB package for implementing COnstraint-based Reconstruction and Analysis methods that were used to simulate, analyze and predict metabolic phenotypes for the *Nannochloropsis* genome-scale model.

In this project, COBRA toolbox 2.0 with several built-in scripts such as gapfind (identifying gaps in the model), BiomassPrecursorcheck (checking if all the precursors for the biomass reaction are synthesized by the model) and OptimizeCbModel (optimizing the flux through the biomass reaction under the constraints given by the rest of the model) were used to reconstruct, refine and simulate metabolic model.

3.1.3 Linear solver

Gurobi5 [59] was used as the linear program solver. It is a solver that can handle all of major problem styles such as LP (Linear Programming), MILP (Mixed Integer Linear Programming), and MIQP (Mixed-Integer Quadratic Programming). The solver is used in many industries with

various applications such as inventory optimization and inventory stocking and reordering. In this project, Gurobi5 was used for the optimization of a linear objective function. The main objective function is the biomass production. The objective function was also changed at some points during the construction process to the production of biomass precursors.

3.1.4 Cytoscape [60]

Cytoscape is an open source software platform for network visualization. It is frequently used by scientists worldwide to analyze biochemical interactions such protein-protein, protein-gene, and gene-gene interactions. The software provides basic tools for layout and analyzing the network.

In this project, Cytoscape was used to visualize primary pathways of the model. The layout was adjusted manually based on pathway schemes from KEGG. Highly connected nodes representing metabolites or cofactors such as, ATP, ADP, H_2O, orthophosphate and diphosphate were removed to allow an easier visualization of the metabolic network.

3.2 Database

Several databases were utilized in order to improve the quality of the *Nannochloropsis* model, both in terms of checking the genomic material for the presence of certain enzymes, and for adding reactions that were confirmed to be present. The online resources that have most frequently been used to reconstruct the model are listed in Table 1 below.

Table 1: Databases list. During the model construction process, NCBI was used to collect the genome information of the *Nannochloropsis.* The draft reconstruction was automatically generated by using PlantSEED and displayed in KBase. Pathway and reaction data in the manual curation process were extracted mainly from KEGG and *Nannochloropsis* genome portal.

Database	Link
Genome database	
The National Center for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/genome/?term=Nannochloropsis
Model platform	
PlantSEED portal	http://bioseed.mcs.anl.gov/~seaver/FIG/seedviewer.cgi?page=PlantSEED
KBase: The Department of Energy Systems Biology Knowledgebase	https://kbase.us/
Pathway database	
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg/pathway.html
Metacyc	http://metacyc.org/
Enzyme database	
UniproKB/Swissprot	http://web.expasy.org/docs/relnotes/relstat.html
Braunschweig Enzyme Database (BRENDA)	http://www.brenda-enzymes.org/
Nannochloropsis genome portal	http://www.nannochloropsis.org/

The National Center for Biotechnology Information (NCBI) [6] is part of the United States National Library of Medicine. The database provides biomedical and genomic data. In this project, genome information of *Nannochloropsis* was retrieved from NCBI.

The PlantSEED [61] original from the SEED (http://pubseed.theseed.org/), a resource for the generation, optimization, curation and analysis of genome scale model with the focus on microorganism, is specially developed for plant. It offers the tool for gene re-annotation and automated reconstruction of the metabolic network. In this work, PlantSEED was used to establish the initial draft reconstruction of *Nannochloropsis*. In PlantSEED, each compartment has different ID (Table 2).

Table 2. PlantSEED ID for each cellular compartment in organism

The cellular compartment	PlantSEED ID	The cellular compartment	PlantSEED ID
The extra cellular environment	e0	The nucleus	n0
The peroxisome	x0	The mitochondria	m0
The cytosol	c0	The endoplasmic reticulum	r0
The chloroplast	d0	The vacuole	v0
The gogli apparatus	g0	The cell wall	w0

The reaction in PlantSEED is started with ‘rxn’ followed by five digit number and compartment ID, for example the pyruvate kinase reaction in the cytosol is named as rxns00148[c0]. The metabolite is identified as ‘cpd’ + five digit number and compartment ID. For instance, water in the cytosol is cpd00001[c0]. Reactions and metabolites in the *Nannochoropsis* model will be named as the same format in plantSEED.

KBase [62] (the Department of Energy Systems Biology Knowledgebase) is a software and data platform that offers the tool to simulate and analyze the metabolic model. The initial draft reconstruction generated from PlantSEED in this project was displayed in KBase. Automatic gap filling was carried out here.

KEGG (Kyoto Encyclopedia of Genes and Genomes) [28] is a database resources that provides information about pathways and reactions in organisms. Added reactions and metabolites in the model will use KEGGID with some modifications (will be explained in the section 3.3.4 Manual refinement of the model). In KEGG, reaction is started with a letter ‘R’ (reaction) and followed by 5 digit numbers. For instance, reaction alpha-D-Glucose 1-phosphate 1,6-phosphomutase is identified as R00959 in KEGG. Similarly, metabolite in the database is started with a letter ‘C’ (compound) and followed by 5 digit numbers. So, D-Glucose 1-phosphate, reactant in the previous reaction is identified as C00103.

Nannochloropsis genome portal [63] offers tool to retrieve annotated gene in the *Nannochloropsis*.

KEGG and *Nannochloropsis* genome portal are the two main databases that provide pathway and reaction data for the manual *Nannochloropsis* model refinement.

Databases such as **Metacyc**, **UniproKB/Swissprot** and **BRENDA** provide information about enzyme, pathway and reaction. These databases were used when pathway and reaction data cannot be found in KEGG.

3.3 Method

The approach utilized for building the *Nannochloropsis* model was based on previous published protocols for reconstruction of metabolic networks [14] [1] [13]. The framework pipeline consists of four main steps to reconstruct the metabolic model for *Nannochloropsis*. These are shown in Figure 17.

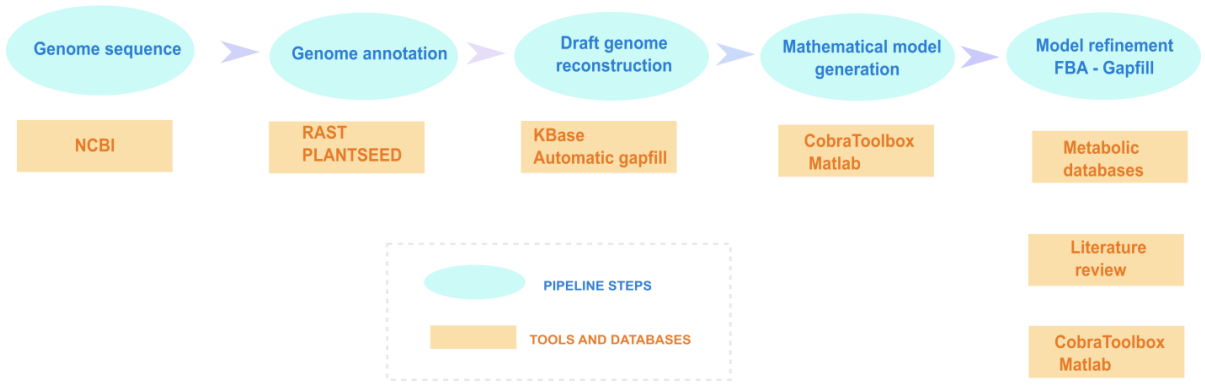


Figure 17. Model reconstruction framework. The reconstruction started by collecting genome information from NCBI. Follow by the use of RAST server and PLANTSEED pipeline to generate metabolic functions for each gene in the genome. A draft genome reconstruction was then created on KBase. Automatic gapfill was also done here. Next, COBRA Toolbox was used to import KBase initial draft model into MATLAB. The model was then refined based on metabolic databases and literature review. Flux-balance analysis (FBA) was carried out with the help of COBRA Toolbox

The reconstruction made in this project started by collecting genome information about *Nannochloropsis* from NCBI, followed by the use of the RAST server and the PlantSEED pipeline to generate metabolic reactions for each gene in the genome that could code for a biologically active protein. A draft genome reconstruction was then created in KBase, which is an online tool meant for presentation and analysis of metabolic networks. Automatic gapfilling was also performed within the KBase framework. The MATLAB format model within KBase was then imported into MATLAB. The model was then refined by utilizing metabolic databases and literature search. Flux-balance analysis (FBA) was carried out by utilizing COBRA toolbox. A detailed description of the steps involved in creating and curating the *Nannochloropsis* model is given below.

3.3.1 Stage 1 – Obtaining a draft reconstruction

The genome of *Nannochloropsis* was retrieved from NCBI. Amongst the four *Nannochloropsis* species available on NCBI, including *N. oceanica*, *N. gaditana*, *N. salina* and *N. limnetica*, the genome annotation of *N. gaditana* was the most complete one. Therefore, the draft reconstruction was built based on the genome annotation for *N. gaditana*.

The available genomic information was used to find open reading frames (ORFs) which predict the possible gene coding sequence in the genome, and subsequently find their corresponding proteins. FASTA sequences of all proteins including hypothetical proteins from the nuclear, the mitochondrial and the chloroplast genomes of *N. gaditana* were extracted from genbank files available from NCBI. The resulting FASTA file was then used by PlantSEED to assign metabolic functions to individual genes in the *N. gaditana* genome. Since there was no available genome information about *N. gaditada* in the databases utilized by PlantSeed, it therefore utilized RAST instead, alongside with the genetic information available in its database for Viridiplantae – a plant that evolved from a green alga - the closest available genome to *N. gaditana*.

The list of metabolic reactions was used to generate a metabolic reconstruction, which was performed automatically by KBase. The list was utilized to assign information about reactants, products, directionality, subsystem of different reactions and cellular location for enzymes. A biomass reaction was constructed up by PlantSEED base on the information for Viridiplantae.

3.3.2 Stage 2 - Curation of the draft reconstruction

In order to make the reconstructed model to produce biomass, an automatic gap-filling (which is the process of either by adding a minimum number of new reactions or by changing the reversibility of the existing reactions to connect the dead-end metabolites to the network) was carried out within KBase. The automatic gap-filling process was based on the metabolism of Viridiplantae.

Eventually, a draft metabolic reconstruction was obtained. This would be further refined and validated.

3.3.3 Stage 3 - Conversion of the genome-scale reconstruction to a MATLAB model

To allow the further refinement of the model, the draft model was extracted from KBase and subsequently imported to into MATLAB. The model was stored in a structure containing several vectors and matrices, the most central one being the stoichiometric matrix, the metabolite vector, and the reaction vector. The relationship between the stoichiometric matrix and the metabolite vector is explained in depth in chapter 2, section 2.2.3.

However, after imported to MATLAB, we noticed some problems with the model that was probably caused by KBase and that needed to be fixed. For example, some reactions contain reactants and products in the cytosol while there reaction IDs are in other compartments. Therefore, the next step was to update the metabolites to their real compartment based on the actual location of the reaction. For instance, the reaction rxn00001_d0 has the corresponding reaction equation cpdxxxx_c0 + cpdxxxx_c0 => cpdxxxx_c0, which means that all the metabolites involved are located in the cytoplasm, despite of the reaction id indicating the reaction takes place in the chloroplast. So the reaction was updated to rxn00001_d0 : cpdxxxx_d0 + cpdxxxx_d0 => cpdxxxx_d0.

Since MATLAB distinguishes between metabolites in different compartments, import reactions are crucial for allowing the movement of metabolites between different compartments. Moreover, the metabolic reconstruction built in KBase was based on plant metabolism, so manual refinement was most certainly required in order to customize the model to *Nannochloropsis* metabolism.

3.3.4 Stage 4 – Manual refinement of the model

Because there are differences between the metabolism in plants and in algae, the draft model produce by PlantSEED contained many redundant reactions, and even compartments. As an example, the initial model included a vacuole compartment. It was removed, because this compartment is not present in algae. Despite of the KBase gap-filling, the first draft also contained a large number of gaps in the network, making many important metabolic pathways incomplete. Hence, the manual curation step required removal of plant-specific reactions, non-existing compartments in algae and the elimination of numerous gaps. Moreover, reactions that involve in secondary metabolites biosynthesis will also be removed. In addition, the draft model

obtained from PlantSEED provided no information about which pathway the reaction participate in. Therefore, in the step of manual curation, pathway information was manually assigned to all reactions.

Gap filling

In the initial model, there were several reactions that produced metabolites that were not connected to any other reaction. Such metabolites are called dead-end metabolites, and they produce gaps in the metabolic network. As mentioned earlier in the section 2.2.3, gaps in metabolic models can be categorized into two groups: non-consumed (RNCs) and non-produced metabolites (RNPs). Non-consumed metabolites block the upstream reactions, while non-produced metabolites block the downstream reactions. Gaps could arise due to the limitation of an automated procedure for retrieving reaction details. In some cases, a reaction may be described in a too general way, which might also create gaps in the model. For example, the reaction driven by hexokinase (EC 2.7.1.1) is defined as " ATP + D-Hexose <=> ADP + D-Hexose 6- phosphate". In this case, hexose is a general name for a 6-carbon monosaccharide, not a name for an actual 6-carbon molecule. As long as the name listed in the model is "hexose" and not for example "glucose", the metabolite created by hexokinase in the reaction described above cannot link to any other reaction in the model. Therefore, it is crucial to customize the reaction to the more *Nannochloropsis*-specific version "ATP + D-Glucose <=> ADP + D-Glucose 6-phosphate" which is also the first step of the glycolysis pathway. In addition, some enzymes can convert multiple substrates. So even though an enzyme could be present in the model catalyzing one reaction, there may still be missing reactions that are catalyzed by the same enzyme in the model, and these would have to be added manually.

A process known as gap filling which is to add one or more reactions that connects a dead-end metabolite with other metabolites in the network was carried out to solve the model's gaps. The reactions that can be added to eliminate model's gaps were categorized into three groups as being displayed in the Figure 18 [64]. Depending on the available of literature, biochemical data or metabolic data, the gaps can be solved by either changing the directionality of the current reaction or adding one or more reactions that allow connecting a dead-end metabolite with other metabolites of the network. These two types of reactions are referred as category I and II (Figure 18) [64]. The third category is to add reaction that link the current gaps with the other metabolites in the network and then export this metabolite to the extracellular environment [64].

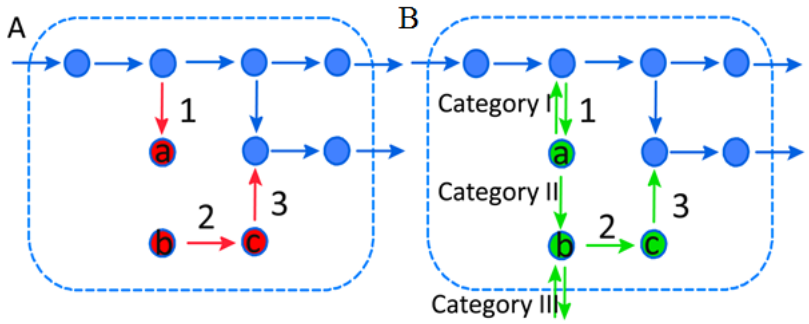


Figure 18. Three categories of the reactions that can be added to eliminate the network gaps. a) The metabolites a and b are gaps, namely RNCs and RNPs, respectively. b) These gaps can be solved by either changing the directionality of the current reaction (category I) or adding one or more reactions that allow connecting a dead-end metabolite with other metabolites of the network (category II). Adding reaction that link the current gaps with the other metabolites in the network and then export this metabolite to the extracellular environment (category III) can also help to remove the gaps. The Figure was obtained from Rolfsson, O., Palsson B.Ø., and Thiele I. [64].

In this case, reactions from KEGG were added to the model after modifying their IDs in KEGG to include the ID of the compartment where the reaction takes place (i.e. R01068_c0, R01068 from the KEGG ID of the reaction in question, fused with _c0 to denote that the reaction takes place in the cytoplasm). New metabolites introduced from KEGG were marked as replacing the 'C' in KEGGID by 'cpd' and adding compartment ID in the suffix. For example, an additional KEGG metabolite C00123 in the cytoplasm would be named as 'cpd00123_c0' in the model. A letter 'K' (KEGG) was adding to the new metabolites that were shown to have overlapping IDs with the existing metabolites in the model. For example, the model already included a metabolite cpd00085 which in the model is Alanine. Hence, if the metabolite having the KEGG ID C00085 (Glutamine) were to be added to the model, it would have to be named cpdK0085, in order to not confuse the two metabolites.

Addition of new reactions may also introduce new properties for the organism in question, so physiological data or evidence that the reaction is present in *Nannochloropsis* should be used to determine whether or not the gaps should be filled [1]. If a reaction without physiological data has to be added in order for the model to run, the reaction should be noted as a hypothesis reaction.

Determination of gaps

Let m (number of rows, corresponding to number of metabolites) and n (number of columns, corresponding to number of reactions) denote the dimension of the stoichiometric matrix. From now on, the metabolite index and reaction index in the stoichiometric matrix will be denoted as i and j, respectively.

As mentioned earlier in the section 2.2.3, the presence of root non consumed (RNCs) and/or root non produced (RNPs) gaps in the model can block all downstream or upstream reactions, respectively. Metabolites involve in these blocked reactions are referred as downstream or upstream gaps. There are several approaches to distinguishing between root-gaps (non-consumed or non-produced metabolites) and downstream or upstream gaps. The root-gap can be determined as a metabolite that all reactions it participated in was belong to blocked reactions [55]. Yet, it cannot tell whether the found root gap is a root non-consumed (RNC) or root non-produced (RNP). However, determining whether a root gap is RNC or RNP is quite simple when considering its role in the set of reaction it participates in. If it is only involved in irreversible reactions as a reactant, then it is a RNP metabolite. Otherwise, if it is a product in all reactions it participates in, it is a RNC metabolite.

Another approach which can be used to find out whether a metabolite is RNC or RNP is to investigate the network topology [1]. In the S matrix, the rows consist of stoichiometric coefficients of different metabolites, while the columns indicate the reactions. For each stoichiometric coefficient a_{ij} in the S-matrix, there are three different options; $a_{ij} < 0$, $a_{ij} > 0$, and $a_{ij} = 0$ which denote that the metabolite either is consumed, produced or does not participate in the reaction in question. This means that most of the matrix will consist of zeroes, and also that, if all of the stoichiometric coefficients a_{ij} belonging to one metabolite are either only positive or only negative, then it can be said to be only consumed (RNP) or only produced (RNC).

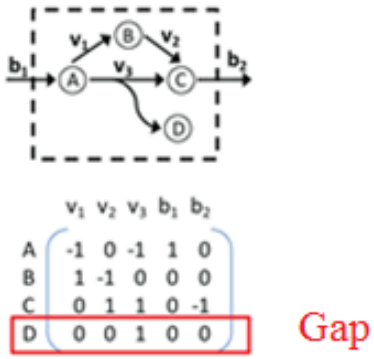


Figure 19. Connectivity gaps in the network. For metabolite D, there are only positive stoichiometric coefficients in the S matrix. This indicates that this particular metabolite is only produced, and not consumed by the model. Hence, this metabolite is a root-non consumed (RNC) metabolite. This in turn indicates a need for more reactions in the network. The figure was obtained from Thiele, I. and B.Ø. Palsson [1].

For instance, as being described in Figure 19, metabolite D is a dead-end metabolite because it only has positive stoichiometric coefficients in the stoichiometric matrix. Since no negative coefficients are found for metabolite D, it means that this compound is only produced in reaction 3 and is not consumed in any reactions. So that makes D a non-consumed metabolite, which will block reaction 3, and make its corresponding flux v_3 zero if optimization is carried out.

However, adding reactions in order to fill a gap, or to remove dead-end metabolites can be a complex process. Unless a reaction is added based on biochemical data or literature, adding new reactions could possibly introduce new properties to the organism, and should therefore be done with caution. Visualization of the metabolic network will help reveal inconsistencies in the network, and will also give understanding of the relationship between gaps blocked reactions, which will be a great advantage when performing gap-filling [55]. To visualize gaps, it is very effective to do it in a pathway-specific manner. In this project, intensive pathway analysis was carried out manually, based on existing pathway schemes retrieved from KEGG in order to investigate the gaps and dead-end metabolites.

Pathway analysis

As already stated, the initial draft only included pathway and subsystem information to a very small extent. Therefore, this information was manually assigned to the model for each separate reaction. All reactions belonging to the same pathway were then extracted and imported into Cytoscape for visualization. Highly connected nodes such as cofactors like hydrogen, ATP, NADH, coenzyme A, oxygen or water were removed to make visualization easier. The layout was adjusted manually based on pathway schemes retrieved from KEGG. Metabolic pathways being essential for cellular function, such as lipid synthesis, glycolysis, the Krebs cycle, the pentose phosphate pathway, amino acid synthesis and nucleotide synthesis were analyzed.

Biomass precursors analysis

Because the flux through the biomass reaction of the initial draft model was zero after it was imported into MATLAB, the biomass precursors were checked to ensure that they were actually produced by the model. The biomass precursors were analyzed followed Thiele, I. and B.Ø. Palsson [1]. Then the missing biomass precursors were added after gathering information about reactions that could be missing. Finding possible missing reactions that could affect the biomass function included both inspecting databases and searching in literature. The gap-filling process were performed according to the method used in Thiele, I. and B.Ø. Palsson (2010) [1] and is explained in Figure 20. A script (provided in the appendix 1) was written based on the scheme to facilitate the gap-filling process.

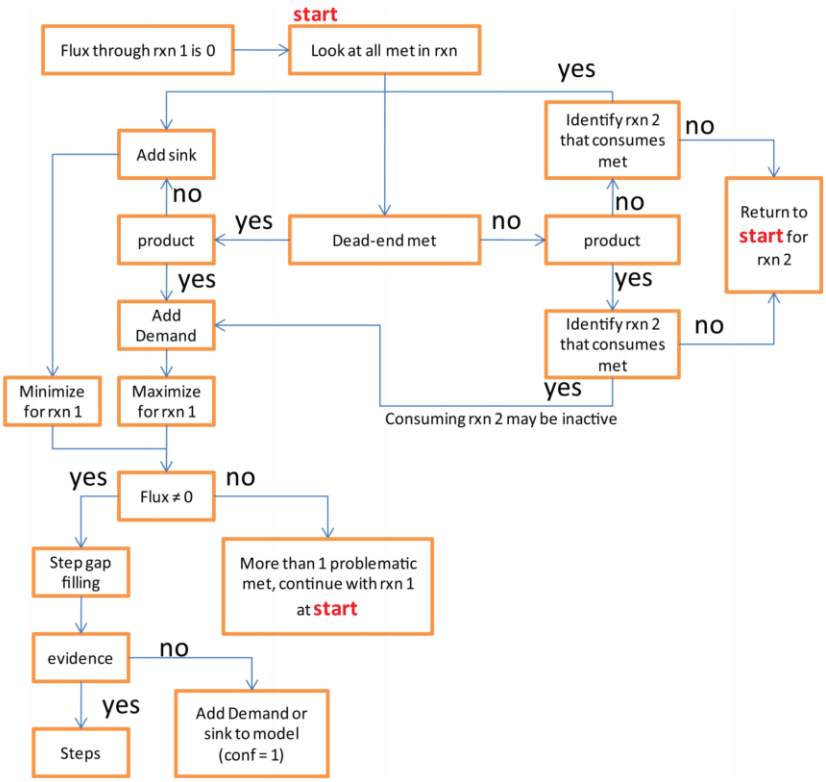


Figure 20. Scheme for gap-filling used to include reactions introduced to correct for missing biomass precursors. If the reaction 1 (rxns 1) that produce biomass precursor carry zero flux, all metabolites participate in that reaction will be checked whether they are dead-end metabolites or not. If yes, and they are the product in the rxns 1, add demand reaction. If they are the reactants, add sink reaction. If the rxns 1 carry flux, find reaction to fill the gap. If the rxns 1 cannot carry flux, it means more than one problematic metabolite in the rxns 1, start again with the rxns 1. In case, no dead-end metabolites in the rxns 1 is found, continue with the rxns 2 that produce the biomass precursors, repeat the same steps as in rxns 1. The figure was obtained from Thiele, I. and B.Ø. Palsson [1].

Steps included in adding reactions for gap filling

All candidate reactions were analysis, and subsequently checked for organism specificity before being added to the model. As a rule of thumb, enzymatic reactions that was added to the model only based on gene-annotation, should receive a low confidence score [1]. Physiological data and evidence indicating a certain reaction should present in the *Nannochloropsis* model was collected from the database *Nannochloropsis* genome PORTAL [63] , while detailed info about reactions was taken from KEGG. Each reaction was assigned to its corresponding subsystem in KEGG. Subsystem information of reactions that occur in several KEGG maps will reflect their primary functions [1]. Intracellular transport reactions were also added to the model to allow connection between compartments in the cell.

4 Results and discussion

The aim of the thesis is to construct a genome scale metabolic model of *Nannochloropsis*. In order to do that, a draft reconstruction created from the alga genome was obtained from PlantSEED. The initial model was imported into MATLAB to facilitate further refinement. Manual curation was carried out to eliminate the network gaps.

The results are presented in five parts. Section 4.1 – section 4.2 describes outcomes of the four stages in the reconstruction process, described in chapters 3.3.1 – 3.3.4 in the materials and methods part. Section 4.3 describes the characteristics of the final model. Section 4.4 and 4.5 reflect the methodology and the findings.

A summary of the results is displayed in Figure 21, where each part will be explained throughout this section.

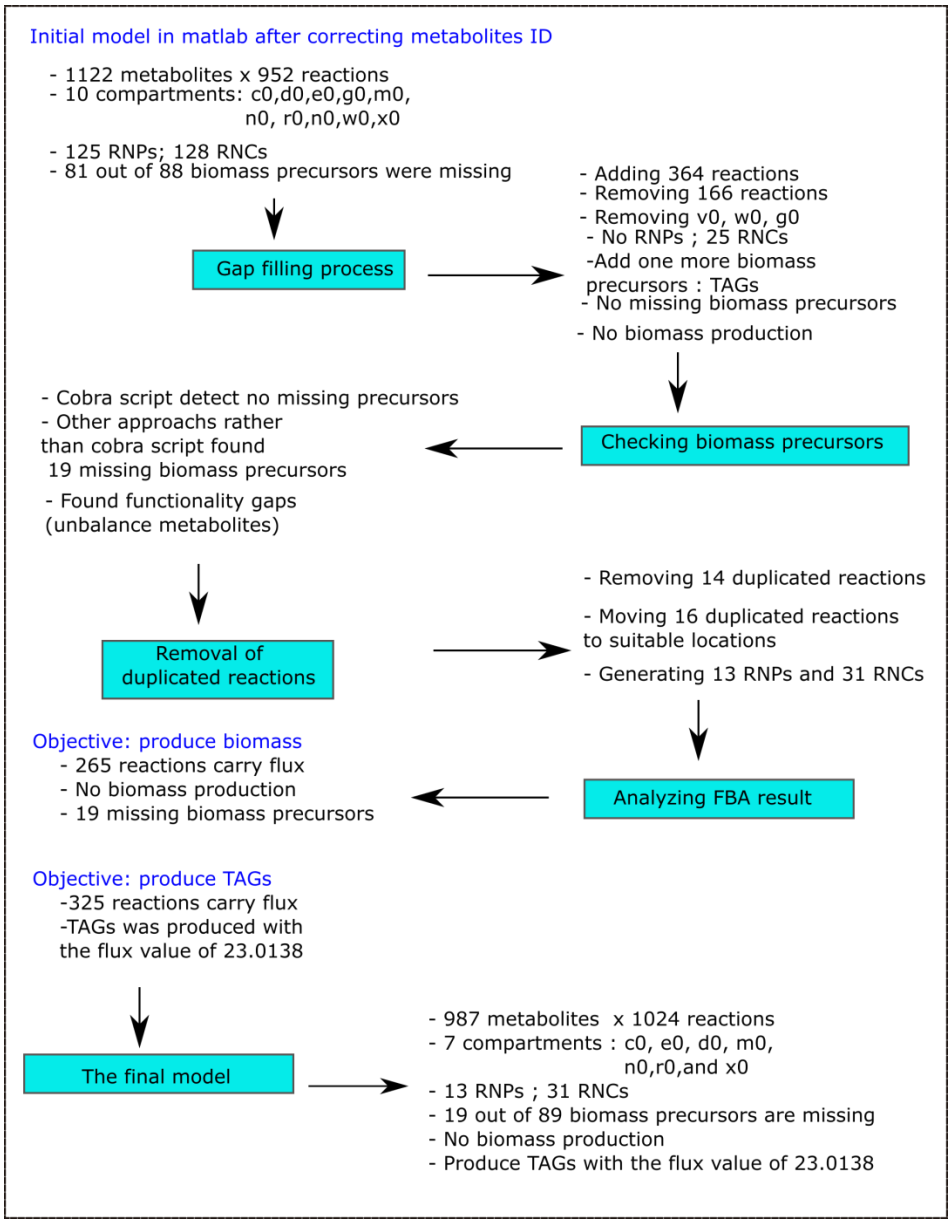


Figure 21. Development of the model according to the steps described in the methods section. In the first step, the initial model was generated with 1122 metabolites and 952 reactions in ten different compartments. The gap filling process eliminated all RNPs and most of the RNCs. According to the COBRA script biomassPrecursorCheck, no missing biomass precursors were found. However, when using the function 'OptimizeCbModel' in COBRA, the biomass reaction had no flux. The following biomass precursor check resulted in 19 biomass precursors that could not be produced. Although the model at its current state is not able to produce biomass, triacylglycerol can be produced at a rate of 23.0138 when the objective function to this metabolites.

4.1 Stage 1-2-3 Obtaining a draft reconstruction

The draft metabolic model of *Nannochloropsis* sp. made using PlantSEED and displayed in KBase included 383 genes and 750 reactions with 814 compounds in 10 compartments. The initial ten compartments were cytosol (c0), the chloroplast (or plastid) (d0), the mitochondria (m0), the endoplasmic reticulum (r0), the peroxisome (x0), the vacuole (v0), cell wall (w0), nucleus (n0), the Golgi apparatus (g0) and the extracellular environment to the cell (e0), which is not really a compartment, but was treated that way in the model for simplicity.

As the objective function of the model was to produce biomass, a biomass reaction was assigned to the model by PlantSEED. This was based on the metabolism of Viridiplantae. Eighty-eight reactants that represent several basic building blocks needed for cell growth were included to simulate biomass production in the algae. The components in the biomass reaction are listed in Table 3. At the current stage, the biomass reaction is based on the biomass reaction of Viridiplantae. Viridiplantae is assumed to be a good approximation, since the main components of *Nannochloropsis* are generally similar to those of plants [33].The draft reconstruction cannot produce biomass due to the presence of gaps.

Table 3. Eighty-eight reactants for biomass reactions in the initial model

Components	Major reactants
Photosynthetic lipids	Zeatin, Beta _carotene, Chlorophyll a, Plastoquinone 9.
Vitamin and cofactors	Folinic acid, Biotin, Ascorbate, Alpha tocopherol Tetrahydrofolate, Formyltetrahydrofolate, Methyltetrahydrofolate, Methylenetetrahydrofolate, Methenyltetrahydrofolate CoA, Thiamin diphosphate, Ethylene, pantetheine 4 phosphate, Heme
Trace elements	Cl, Potassium cation
Lipid	Hexadecanoic acid, Octadecanoic acid, Octadecanoic_acid. Beta_Sitosterol, Campesterol, Stigmasterol Beta_D_Glucopyranuronic_acid, R__Lipoic_acid
Carbohydrates and sugars	beta_D_Ribofuranose, alpha_D_Glucose, alpha_D_Mannose, beta_D_Fructose, alpha_D_Galactose, xylose, Alpha rhamnose, Alpha arabinose, Glycerol 3 phosphate
Nucleotides and energy molecule	ATP, NAD,NADH, NADPH, NADP, FAD GTP,FMN, CTP, UTP, dATP, dGTP, dCTP, dTTP UDP_6_sulfoquinovose
Amino acids	Alanine, Glutamate, Glycine, Lysine, Aspartate, Arginine, Glutamine, Serine, Methionine, Tryptophan, Phenylalanine, Tyrosine, Cysteine, Leucine, Histidine, Proline, Asparagine, Valine, Threonine, Isoleucine
Others	H2O, S Adenosyl L methionine, Oxaloacetate, Choline, myo Inositol, S Malate, Citrate, S lactate, Ethanolamine, D galacturonate, cis Aconitate, 4 Coumarate, Indole 3 acetate, Ferulate, Luetin, Phytosphingosine, Pyridoxal phosphate

The automatic gap-filling process performed by KBase increased the number of reactions from 750 to 952 reactions and from 814 to 903 compounds. The number of compartments and genes remained unchanged. However, the model produced biomass when it was still in KBase with the flux value of 22.6689 with all reactions carrying flux.

The draft reconstruction from KBase was then formulated to a MATLAB model using COBRA Toolbox.

As mentioned in the section 3.3.3, after importing to MATLAB, the model contains some reactions, 161 reactions to be specific, contain reactants and products in the cytosol while there reaction IDs are in other compartments. Hence, these reactions were renamed to correct the metabolites compartments identifiers. This yielded a model with 952 reactions and 1122 compounds. At this stage, the model contained more compounds than reactions. More

constraints than fluxes created an over-determined system, whereas the most common situation for metabolic models is underdetermined systems.

The initial mathematical model contained around 125 non-produced metabolites (RNPs) and 128 non-consumed metabolites (RNCs), which yielded 253 gaps in total. As a result, 88% of the reactions were blocked, and only 138 out of 1122 reactions were able to carry flux. At this stage, only 9 out of 88 biomass precursors were produced due to the presence of many incomplete essential pathways.

4.2 Stage 4 – Manual refinement of the model

4.2.1 Gap filling

The presence of gaps (caused by RNPs and RNCs) caused various blocked reactions in all compartments (an overview is given in Figure 22). Out of ten compartments in the model, cytosol was the compartment that contained the largest number of gaps with 77 non-produced and non-consumed metabolites. This accounted for 12% of the total amount of metabolites in this compartment. 89 % of the reactions in cytosol were blocked. The situation in the plastid was similar, with 65 non-produced and non-consumed metabolites, blocking 35% of the metabolites in this compartment. 102 out of 117 reactions were unable to carry flux. 50% of metabolites in the endoplasmic reticulum (ER) were gaps, causing 76% of the reactions in the ER to be blocked. Most of the metabolites in the other compartments such as the golgi apparatus, the vacuole, the peroxisome and the cell wall also created gaps, and consequently blocked most of the reactions in these compartments.

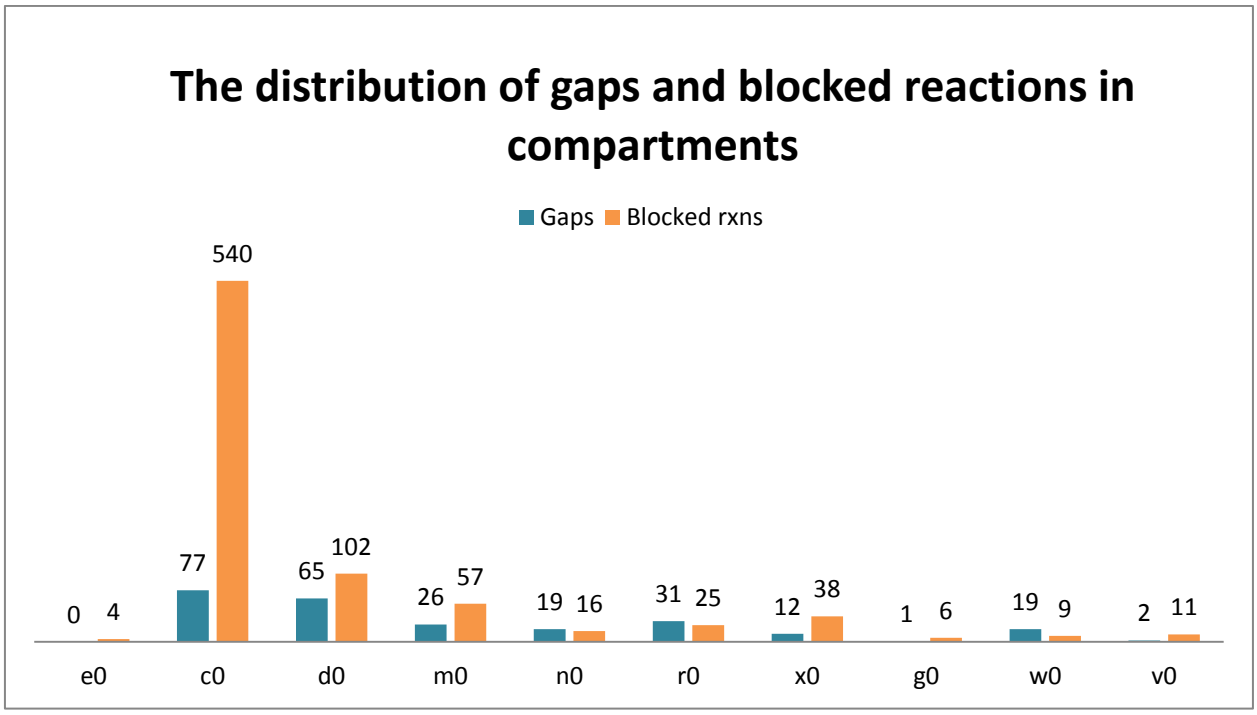


Figure 22. The distribution of gaps and blocked reactions in the different compartments. e0 - extracellular compartment, c0 - cytosol, d0 - plastid, m0 - mitochondria, n0 - nucleus, r0 - endoplasmic reticulum, x0 - peroxisome, g0 - Golgi apparatus, w0 - cell wall, v0 –vacuole

The gaps appeared that they can be grouped by using the pathways they interfere with. Some RNPs and RNCs are listed in the Table 4. As can be seen from this Table, myristoyl ACPs, Decanoyl acp, and Palmitoleoyl ACPs are root gaps and they are all participated in the fatty acid biosynthesis. Similarly, NADPH, NADP and fumarate are root gaps that can be grouped into the Citrate cycle. Hence, for an efficient gap-filling, the primary metabolism of the model was completed and analyzed for its ability to synthesize amino acids, carbohydrates, nucleotides and lipids.

Table 4. The potential pathways that gaps metabolites (RNCs and RNPs) interface with

RNCs	RNPs	Potential pathways
Myristoyl ACPs in the mitochondria	Decanoyl acp in the cytosol	Fatty acid biosynthesis
Palmitoleoyl ACPs	Myristoyl_ACPs_C14H27OSR_c0	Fatty acid biosynthesis
NADPH	NADP__C21H25N7O17P3_m0	Citrate cycle
Fumarate_C4H2O4_m0		Citrate cycle

At the early stages of the model, most of the pathways, including the Calvin–Benson–Bassham cycle (Calvin cycle), the citric acid or tricarboxylic acid (TCA) cycle, glycolysis, and the pentose phosphate pathway were incomplete. However, it is generally accepted that the essential core metabolic pathways are common between most organisms, and especially between plants and algae. This is supported by Vieler, A. and coworkers [21] The *Nannochloropsis* genome contained all the genes coding for proteins involved in the central metabolism. But several of the reactions catalyzed by these enzymes were still missing from the model. Therefore, reactions were added to complete these pathways. In addition to completing these main pathways, reactions were also added to eliminate the gaps in other pathways. However, I have not investigated these pathways in depth yet. 364 reactions were added to fill all gaps in the model. As being displayed in Figure 23, 31% of the total amount of added reactions was involved in lipid synthesis. The other 18% was added to fill gaps in amino acid synthesis. Nucleotide synthesis, and vitamin and cofactor metabolism, counted for other 31% of the adding reactions.

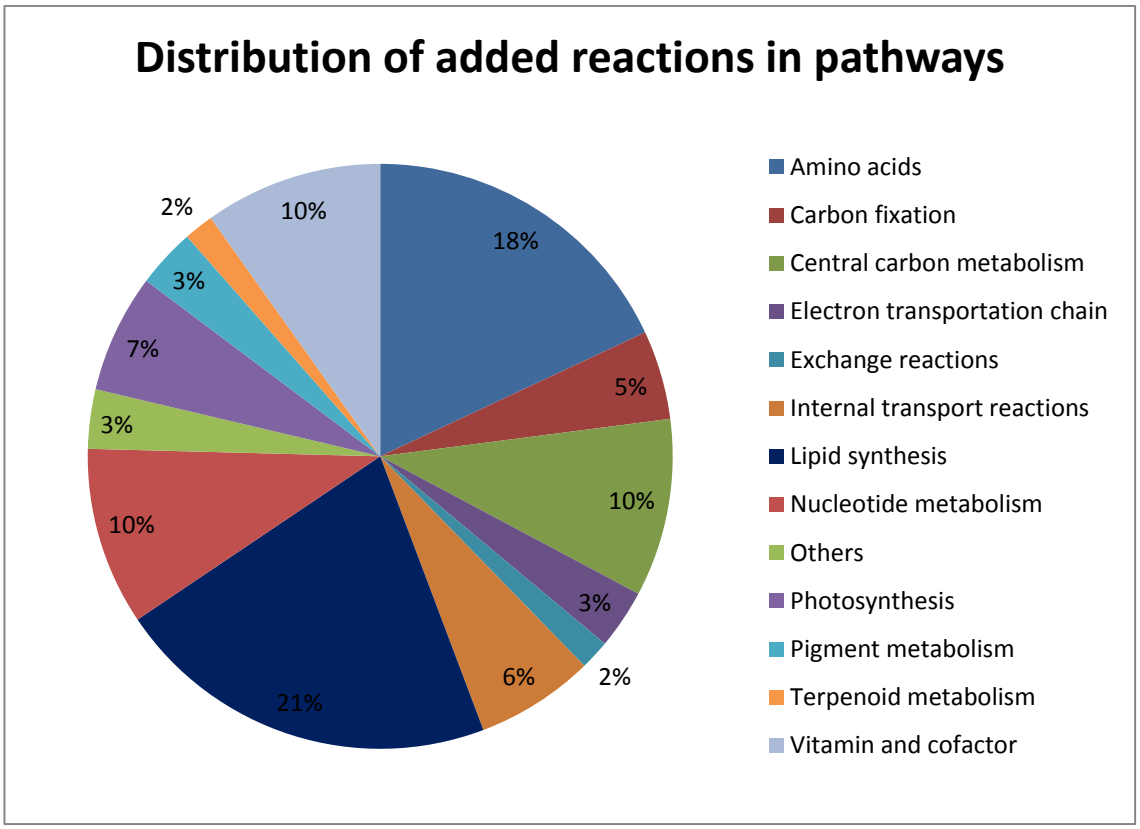


Figure 23. The distribution of the added reactions in pathways. Lipid metabolism contained the highest number of added reactions, followed by the number of reactions in the amino acids biosynthesis. The same number of reactions was added for the nucleotide metabolism and the vitamin and cofactor metabolism.

The central metabolic pathways contained 26 RNCs and 23 RNPs (Figure 24). Six reactions in category I (reactions where the directionality of the reaction in question was reversed) and 97 reactions in category II (new reactions added to fill the gaps in the network) were added to eliminate gaps. No reaction in category III (exchange reaction that allow the export of the gap metabolite to the extra environment) was added. The number of gaps and reactions in category I and II added in each pathway is displayed in Figure 24.

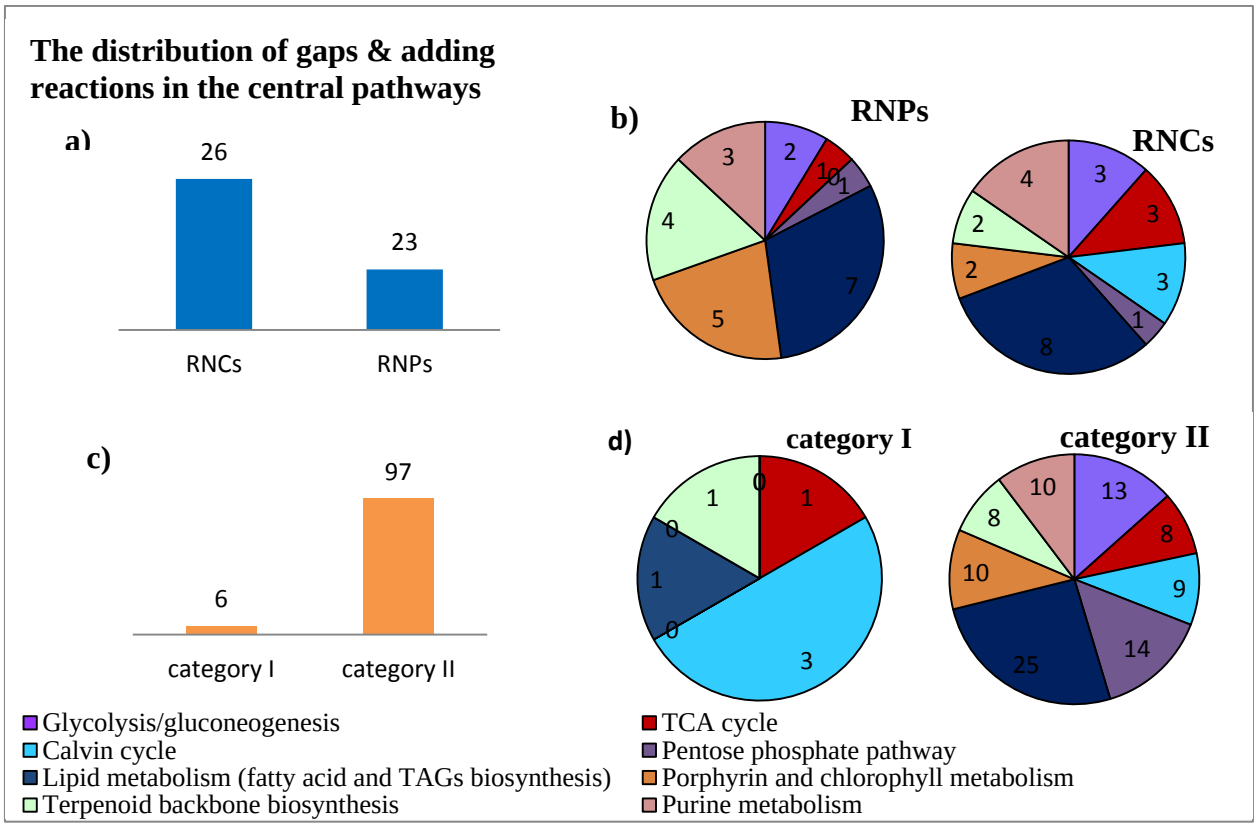


Figure 24. The distribution of gaps and added reactions in central metabolic pathways. a) 26 metabolites in the central pathways were only consumed and not produced (RNPs) in any reaction, while 23 metabolites were only produced and not consumed (RNCs). b) All the analyzed pathways contained RNCs gaps, while RNPs were present in all pathways except the Calvin cycle. c) To eliminate these gaps, the reaction directionality was changed for 6 reactions (category I) and 97 reactions in category II were added. d) Out of six reactions in category I, three of them were added in the Calvin cycle. The other three reactions were added for the TCA cycle, terpenoid backbone biosynthesis and de novo lipid synthesis. The highest numbers of reactions in category II were added to the lipid metabolism, followed by the pentose phosphate pathway, the glycolysis, purine metabolism and the pentose phosphate pathway.

The Calvin cycle also known as the reductive pentose phosphate cycle, is a collection of light-independent biochemical reactions in plant and other photosynthetic organisms. The cycle captures carbon dioxide in order to produce carbohydrates in the chloroplast. In *Nannochloropsis*, all enzymes that participate in this pathway are present in multiple copies [21]. The initial model did not support the simulation of the Calvin cycle because, six reactions (reactions driven by sedoheptulose-bisphosphatase (R01845), ribose-5-phosphate isomerase (R01056), transketolase (R01641), phosphoglycerate kinase (R01512), ribulose-phosphate 3-epimerase (R01529), and D-Ribulose 1,5-bisphosphate oxygenase (R03140)) were initially

located in the incorrect compartment (cytosol). The directionality of three present reactions in the plastid (ribulose-bisphosphate carboxylase (R00024), glyceraldehyde-3-phosphate dehydrogenase (R01061) and fructose-bisphosphate aldolase (R01068)), did not support carrying flux, as shown in Figure 5. There were three missing reactions (phosphoribulokinase (R01523) fructose-bisphosphatase (R00762), and D-Fructose 6-phosphate:D-glyceraldehyde-3-phosphate glycolaldehyde transferase (R01067)) (Figure 25). Therefore, three existing reactions with opposite directionalities were replaced by the new reactions from KEGG. Because the reactions located in cytosol interfered not only with the Calvin cycle, but also with other pathways, such as R01512 and R03140 belonging to glycolysis, the Calvin cycle reactions in cytosol were not removed, but the same reactions were added also to the plastid. Three missing reactions were also included in the now plastid-localized cycle. As a result, thirteen reactions were added to complete the Calvin cycle.

Additional reactions eliminated three RNC metabolites in the pathway. Instead of consuming D_ribulose_1,5_biphosphate, reaction R00024 in the initial model utilized 3_phospho_D_glycerate to produce this metabolite and CO_2. The directionality of this reaction makes these two metabolites RNCs in the plastid. Reversing reaction R00024 eliminated these two RNCs (Figure 5). The incomplete Calvin cycle also created a new RNC, namely glycerone phosphate, which was removed by the addition of a consumption reaction R01829 (Figure 25).

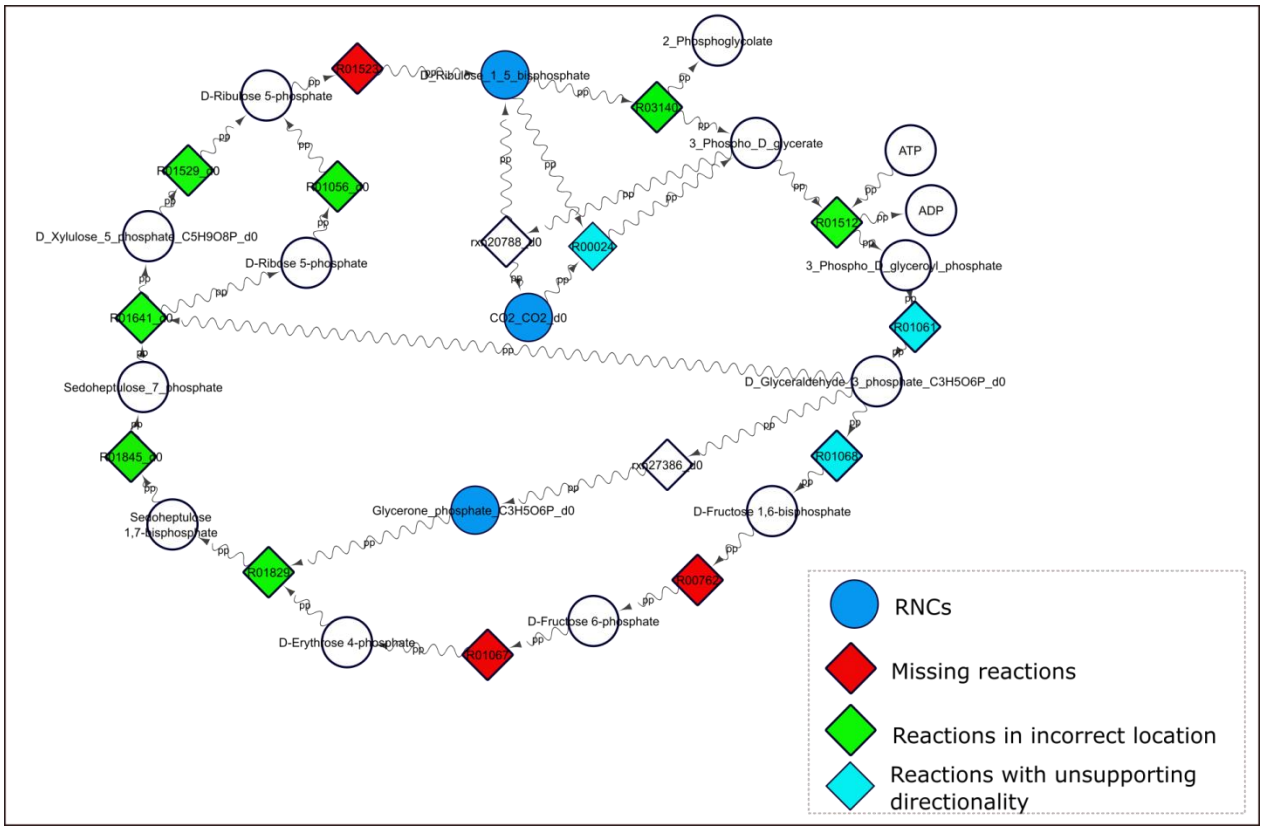


Figure 25. The distribution of RNCs, RNPs and missing reactions in the Calvin cycle. The Calvin cycle contained no RNPs, but three RNPs because the pathway had three missing reactions, seven reactions located in an incorrect compartment and one reaction with a wrong directionality.

Glycolysis is a pathway that all organisms use to produce ATP through the conversion of one molecule of glucose into two molecules of pyruvate [25]. Unlike other organisms, in Nannochloropsis, glycolysis is found in both cytosol and in the plastid, and the enzymes involved in the pathway are coded in multiple gene paralogues in the genome [21]. In our initial model, neither of the glycolysis pathways in cytosol or plastid was completed. The pathway in cytosol was missing one fructose-bisphosphate aldolase reaction, converting D-fructose 1,6-bisphosphate into glycerone phosphate and D-glyceraldehyde 3-phosphate. The pathway in the plastid was more incomplete than that in the cytosol, with eight missing reactions including the ones driven by hexokinase, pyruvate kinase, pyruvate synthase, glucose-6-phosphate isomerase, 6-phosphofructokinase, fructose bisphosphatase, ADP specific phosphofructokinase, and phosphoglycerate mutase. These missing reactions formed one RNCs and one RNPs in the cytosol, whereas, in the plastid there were two RNCs and one RNPs. As being displayed in Figure 26, the directionality of reaction rxn00543_c0 did not support production of ethanol and the consumption of acetaldehyde which make these metabolites RNPs and RNCs, respectively. To eliminate the gaps in the cytosol pathway, the directionality of reaction rxn00543_c0 was changed from

irreversible to reversible. Two reactions, R01518_d0 and R00200_d0 were added to consume the two RNCs 3_phospho_D_glyceral phosphate and phosphoenolpyruvate. However, the directionality of reaction R01518_d0 supports the Calvin cycle rather than the glycolysis pathway. Since the reaction is reversible, the model runs the reaction in the opposite direction to unblock the pathway. The RNP beta_D_Fructose_1,6_biphosphate was only used in reaction rxn19683_c0, but was not produced in any reaction. Three reactions from KEGG were added to produce this RNP in the plastid. The other metabolites, such as pyruvate and 2_phospho_D_glycearate, were neither RNCs nor RNPs,. However, these were produced by reactions in other pathways than glycolysis. Therefore, the missing glycolytic reactions were added. In conclusion, in order to allow the model to produce acetyl-CoA, an important metabolite that also participates in other pathways such as the TCA cycle, 13 reactions were added to the glycolysis pathway taking place in the plastid (Figure 26).

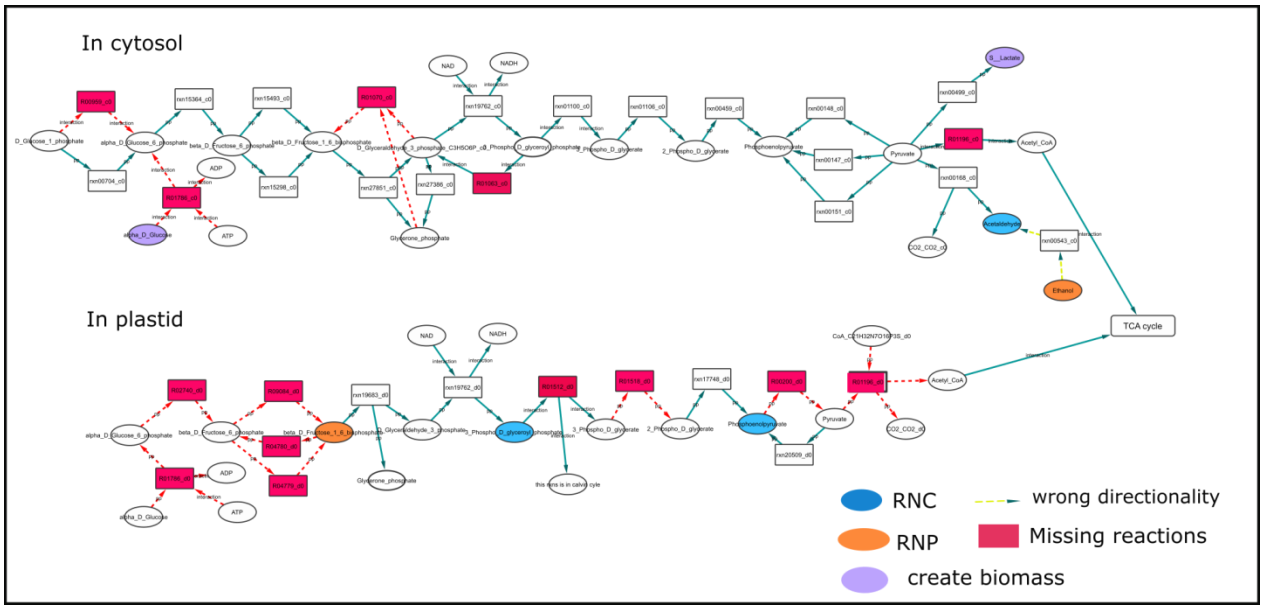


Figure 26. Glycolysis. In *Nannochloropsis*, glycolysis occurs both in the cytosol and in the plastid. None of two versions of the pathway in the initial model were complete. The glycolysis in the plastid contains two RNPs and one RNCs gaps, while that pathway in the cytosol contains one RNPs and one RNCs. This was caused by the lack of 14 reactions and the incorrect directionality of one reaction.

The citric acid cycle (TCA): is among the most basic and essential metabolic pathways, and are utilized by almost all aerobic organisms in order to generate energy [24]. The TCA cycle were amongst the most incomplete pathways in the initial model. The incomplete TCA cycle prevented the model from producing oxoglutarate, a metabolite that also participates in other pathways. The pathway was completed using information provided by Vieler, A. and coworkers (2012) [21] *Nannochloropsis* has multiple paralogues of several enzymes involved in the citric

acid cycle. Therefore, the gaps in the TCA cycle could be eliminated. The initial model was missing the reactions driven by aconitase (R01325), isocitrate dehydrogenase (R01899), pyruvate dehydrogenase (NADP+) (R00210) and ATP citrate synthase (rxn00257_m0). Only four enzymes (2-Oxoglutarate dehydrogenase complex (rxn08094), Succinate:CoA ligase (ADP-forming) (rxn00285), malate dehydrogenase (rxn20068) and (S)-dihydroorotate: fumarate oxidoreductase (rxn01360)) were found in the model, and these were located in the cytosol (Figure 27). However, Vieler, A. and coworkers [21] found only a single copy of aconitase and isocitrate dehydrogenase, and these are predicted to be in mitochondria. Hence, the completed TCA cycle was assigned to be mitochondrial. The citric acid cycle in the mitochondria was initially almost non-present with only the malic enzyme (rxn27459_m0) and fumarate hydratase (rxn19740_m0) present. And in addition, the directionality of the reactions driven by these enzymes did not support flux through the TCA cycle. The directionality of reaction rxn19740_m0 generated two root gaps, namely the RNP S-malate, and the RNC Fumarate. The RNP S-malate was only consumed in reactions rxn19740_m0 and rxn27459_m0, and the RNC Fumarate was only produced in reaction rxn19740_m0.

To complete the pathway, the directionality of reaction rxn19740_m0 was reversed. The TCA cycle reactions in the cytosol plus the missing reactions were added to the mitochondria. In total, eight reactions were added to mitochondria in order to complete and unblock the citrate cycle.

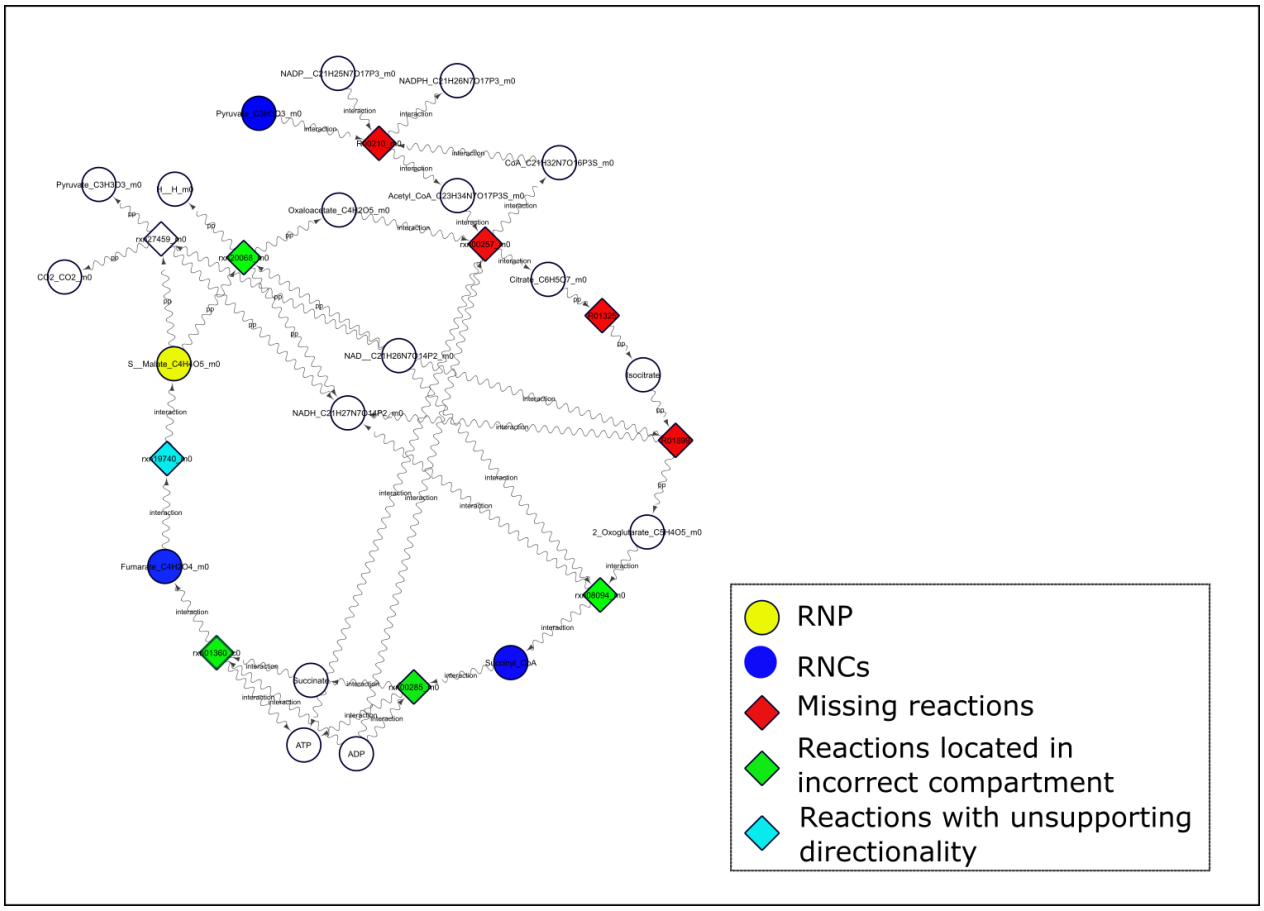


Figure 27. The TCA cycle. Initially, most of the reactions in the TCA cycle were missing. Three reactions were present, but in an incorrect compartment. The TCA cycle initially included three RNCs and one RNP.

The pentose phosphate pathway: The pentose phosphate pathway produces energy in the form of NADPH, and in addition, it produces five-carbons sugars. It also produces important precursors for other pathways, such as Ribose-5-phosphate. The *Nannochloropsis* genome encodes all necessary enzymes in the pentose phosphate pathway [21]. However, only a single copy of ribose-5-phosphate isomerase was found and restricted to be in the plastid [21]. In this regard, a complete pentose phosphate pathway in the model would be assigned in the plastid.

The initial model lacked more than half of the reactions within the pathway, and 14 reactions were added to complete the it (Figure 28). The incomplete pentose phosphate pathway formed one RNC, D_glucono_1,5_lactone_6_phosphate, and one RNP, 6_phospho_D_gluconate (Figure 28). One additional reaction R02035 allowed the consumption of the RNC as well as the production of the RNP.

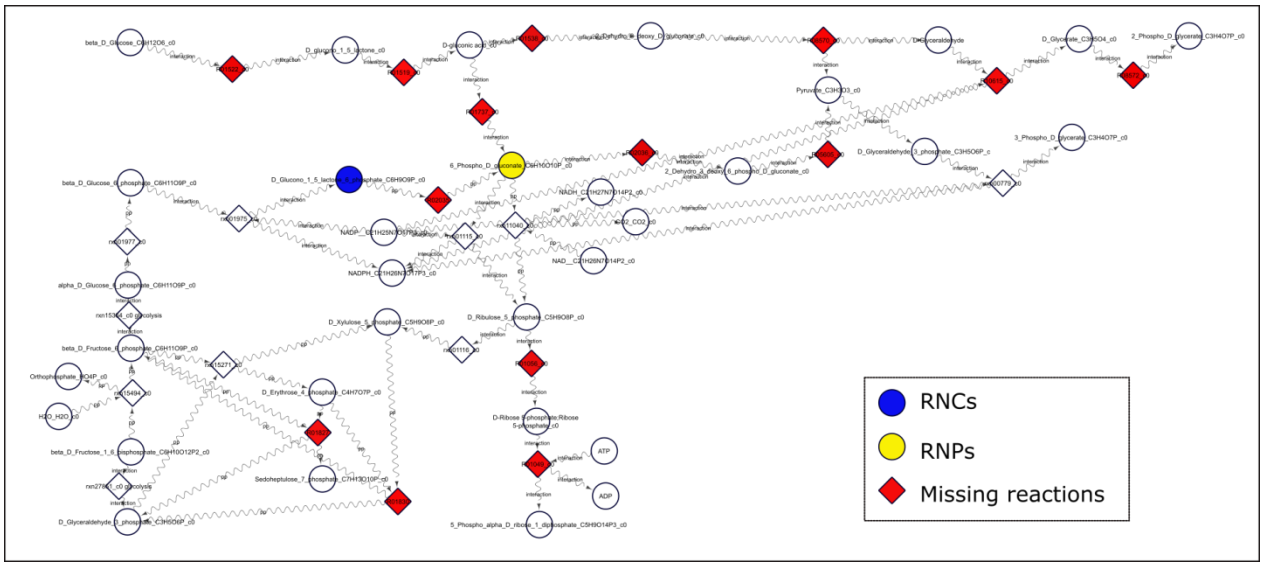


Figure 28. The pentose phosphate pathway. The model initially lacked 14 reactions in this pathway and this caused the formation one RNCs and one RNPs.

Lipid biosynthesis: *Nannochloropsis* is well-known for its ability to produce high amounts of triacylglycerol (TAGs) and polyunsaturated fatty acids such as eicosapentaenoic acid [65]. The biosynthesis of lipids in most plants and algae requires enzyme-catalyzed reactions that are present in different cellular compartments (cytosol, chloroplast, and peroxisome) [65], *Nannochloropsis* contains all the essential enzymes in the central fatty acid biosynthesis in the plastid. However, biosynthetic pathway components in the cytosol and the peroxisome were only incompletely present in the initial model. In order to establish a functional preliminary version of fatty acids biosynthesis, the separate pathways in these two compartments were moved to the plastid and 9 missing reactions were also added to complete the pathway. In addition, the longest chain fatty acids found in *Nannochloropsis is* C20, thereby; reactions that contain longer chain fatty acid such as C22 were removed.

Study shows that the products from fatty acid biosynthesis in form of acyl-CoA is sequentially incorporated into Glycerol 3 phosphate to produce triacylglycerol (TAGs) [66], energy storage molecules in plants, animals and algae [67]. In *Nannochloropsis*, starch synthesis is disappeared and the alga uses TAGs as a main energy source [68]. The fatty acids compositions in *Nannochloropsis* are mostly long chain acyl group with lower desaturation C14:0, C16:0, C16:1, C18:0, C18:1, C18:2, C20:4, C20:5 [19]. The TAGs synthesis pathway, also referred as Kennedy pathway occurs in the plastid and the endoplasmic reticulum of the alga [21]. However, enzymatic reactions for TAGs synthesis, reported by Vieler A. and coworkers (2012) were not present in the initial model. The TAGS biosynthesis reconstruction carried out with Gunvor Røkke, added 17 reactions in the alga lipid synthesis reported by

Vieler A. and coworkers [68] and Coleman, R.A. and Lee D.P. [66] to provide the model ability to simulate the production of TAGs in both plastid and ER compartments. These reactions include reactions that convert normal fatty acids from the fatty acid biosynthesis pathway into fatty acid in form of acyl coA; as well as transport reactions that allow the movements of these fatty acid between the plastid and the endoplasmic reticulum. Artificial reactions that contain eight fatty acids and glycerol that make up TAGs were also added to the plastid and the endoplasmic reticulum. The stoichiometric coefficient for each reactant was formulated according to the published concentration of each component in TAGs [19]. These reactants and their corresponding stoichiometric coefficients in the model are presented in the Table 5.

Table 5. The reactants and their corresponding stoichiometric coefficients that make up TAGs in the model. According to Liu, B. and coworkers [19], the main components of TAGs in Nannochloropsis are mostly long chain acyl group with lower desaturation C14:0, C16:0, C16:1, C18:0, C18:1, C18:2, C20:4, C20:5. Artificial reactions contain these reactants were added to the plastid and the endoplasmic reticulum.

Reactants	Stoichiometric coefficient
C14:0 Tetradecanoyl CoA	7
C16:0 Palmitoyl CoA	49
C16:1 Palmitoleoyl CoA	28
C18:0 Stearoyl CoA	2
C18:1 Oleoyl CoA	7
C18:2 Linoleoyl CoA	1
C20:4 Icosatetraenoyl CoA	2
C20:5 Icosapentaenoyl CoA	3
Glycerol	33

At this stage, glycerol was not produced in the model, to allow the production of TAGs, an exchange reaction was added. This reaction import glycerol from the environment to the plastid and the endoplasmic reticulum where it is used to produce TAGs.

TAGs in the plastid and endoplasmic reticulum were the dead-end products or RNCs which mean they are not consumed in any reactions. In this case, none of the reactions in the lipid metabolism will carry flux because there are dead-ends in the network. TAGs were predicted to form droplet in the cytosol and contribute up to 60% of the biomass of the algae [69]. Hence, to facilitate the flux through this pathway, TAGs in the endoplasmic reticulum and the plastid were transport to the cytosol, then the TAGs in the cytosol was included in the biomass reactions. These reactions are described as follow:

$$TAGs\_d0 + TAGs\_r0 \Rightarrow TAGs\_c0$$

$$TAGs\_c0 \Rightarrow biomass$$

The complete fatty acid and TAGs synthesis pathway eliminated 8 RNCs and 7 RNPs gaps (Figure 9).

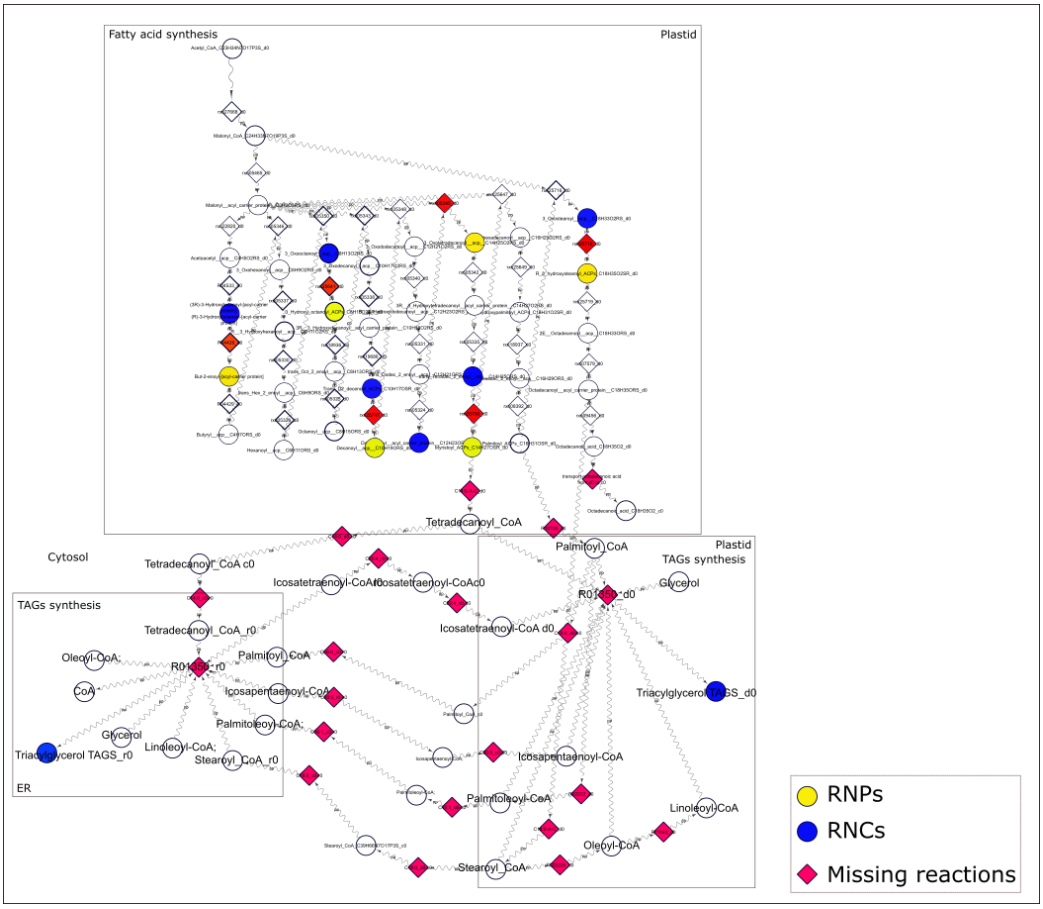


Figure 29. The fatty acid and TAGs biosynthesis. In Nannochloropsis, fatty acid is synthesized in the plastid, while TAGs is produced in the plastid and also in the endoplasmic reticulum membrane. The initial model lacked nine reactions in the fatty acid biosynthesis pathway. All reactions in the TAGs production were missing. As a result, there were eight RNCs and six RNPs in these biosynthesis pathways. Reactions were added to remove these gap metabolites.

Amino acid biosynthesis: According to Vieler and coworkers (2012) [21] de novo amino acid biosynthesis occurs in most bacteria, archaea, fungi, algae, and plants. *Nannochloropsis* is thought to synthesize all essential amino acids [21] [33]. Many of the RNCs and RNPs in the initial model belonged to the amino acids biosynthesis. Hence, all pathways that involved in amino acids synthesis were completed by addition of missing reactions based on KEGG pathway schemes. At this stage, metabolic pathways for each amino acid have not yet been investigated. Instead, reactions were added to eliminate the gaps.

Nucleotide metabolism: Purines and pyrimidines are the two main components in nucleotides, which are both essential components in energy carriers, nucleic acids, and precursors in the synthesis of cofactors such as NADH [70]. The synthesis pathways for nucleotides have been found to be the similar in plants, microorganisms and animals [70]. Nucleotides can be synthesized using 5-phosphoribosyl-1-pyrophosphate (PRPP) in the *de novo* pathway or the salvage pathway [70]. *Nannochloropsis* was shown to have essential genes both for purine and pyrimidine metabolism [33]. However, the initial model lacked complete pathways for both purines and pyrimidines. Hence, the missing reactions were added to complete the pathways. At this stage, both pathways have been assigned to the cytosolic compartment.

Purine metabolism: The lack of 10 reactions in purine synthesis in the initial model created 4 RNCs and 3 RNPs (Figure 30). The necessary reactions were therefore added to eliminate the gaps.

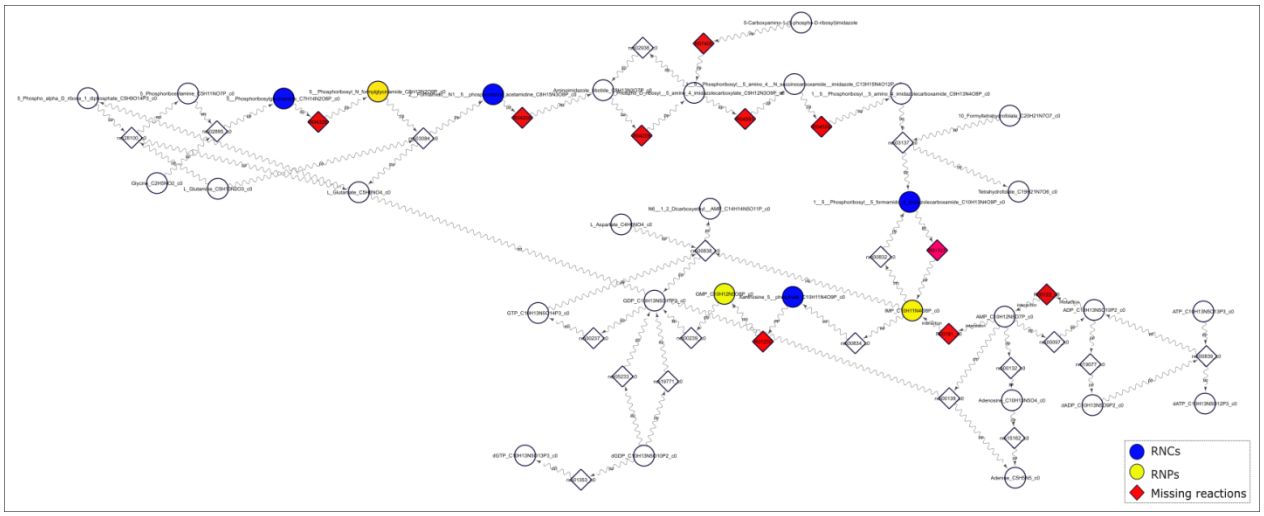


Figure 30. The initial model lacked 10 reactions in the purine metabolism. This caused the existence of four RNCs and three RNPs.

Pyrimidine metabolism: The pyrimidine metabolism in the initial model was incomplete, although no RNCs and RNPs were found. As being displayed in Figure 31, when considering pyrimidine metabolism alone, there are two RNCs and one RNP in the pathway, but these metabolites were not really RNCs or RNPs as they were produced and consumed, respectively, in the reactions of other pathways. For example, S-dihydrooxotate, was produced in a reaction in pyrimidine metabolism, but was only consumed in the reaction Fumarate + 1 S__Dihydroorotate -> 1 Succinate + 1 Orotate belonging to the TAC cycle. *Nannochloropsis* is

predicted to contain all the essential genes for the nucleotide metabolism. Therefore, even though no real RNCs and RNPs were found in the pathway, twenty missing reactions were added.

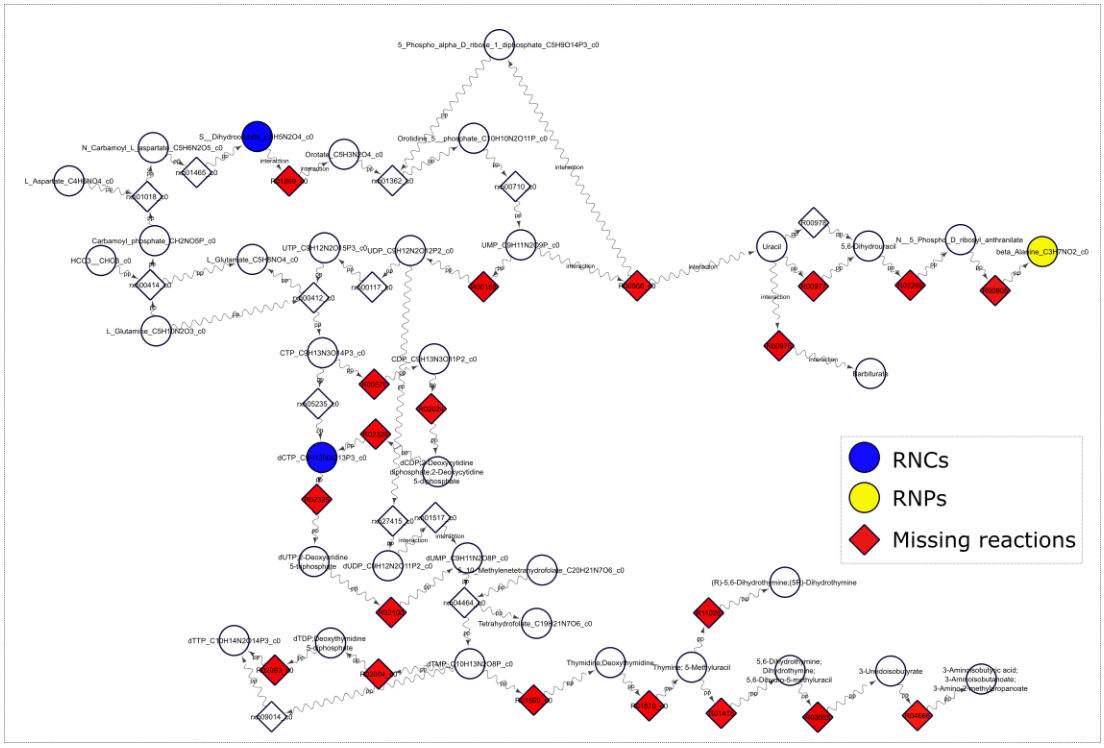


Figure 31. The pyrimidine metabolism. Initially, there were nineteen missing reactions in the pathway. The incomplete pathway resulted in two partial RNPs and RNCs. The missing reactions were added to complete the pathway.

Removal of non-metabolic reactions

The process of gap filling did not only result in the addition of reactions. Many reactions were also removed. Most of these reactions were involved in the biosynthesis of secondary metabolites, which is not covered in the model at this stage (Table 6). In addition, reactions in the vacuole, the cell wall and the Golgi apparatus were removed to eliminate gaps. Moreover, reactions in primary metabolic pathways located in wrong compartment (for instance reactions in the fatty acid biosynthesis located in the cytoplasm instead of the plastid), were also removed. In total, 166 reactions were removed (full list is included in Appendix 7).

Table 6. Removed reactions during the gap filling process. Reactions that interfere with secondary metabolites, or reactions located in an incorrect compartment were removed to eliminate the gaps.

Dead-end metabolite	Reactions	Reason for removal
Dihomomethionine	2 NADPH + 2 Oxygen + 2 H + 1 Dihomomethionine -> 3 H2O + 2 NADP + 1 CO2 + 1 5 Methylthiopentanaldoxime	These reactions interface with the glucosinolate biosynthesis. Glucosinolates are sulfur-rich secondary metabolites that are not covered in the model at this stage
Trihomomethionine	2 NADPH + 2 Oxygen + 2 H + 1 Trihomomethionine -> 3 H2O + 2 NADP + 1 CO2 + 1 6 Methylthiohexanaldoxime	
Tetrahomomethionine	2 NADPH + 2 Oxygen + 2 H + 1 Tetrahomomethionine -> 3 H2O + 2 NADP + 1 CO2 + 1 7 Methylthioheptanaldoxime	
Docosanoic_acid	1 ATP + 1 CoA + 1 Docosanoic acid -> 1 Diphosphate + 1 AMP + 1 H + 1 Docosanoyl CoA	In fatty acid biosynthesis. However, It has been shown that the longest fatty acid in Nannochloropsis is C20. So reactions that include C22 Docosanoic acid are removed
Palmitoyl_CoA	1 Palmitoyl CoA + 1 Acyl sn glycerol 3 phosphate -> 1 CoA + 1 1 9Z octadecenoyl 2 hexadecanoyl sn glycerol 3 phosphate	Fatty acid biosynthesis. However, this reaction occurs in incorrect compartment (cell wall)

The ability to produce biomass precursors was monitored continuously throughout the gap filling process. When all the biomass precursors were produced, the gap-filling process was stopped. After gap-filling was completed, only six missing biomass precursors were found by COBRA (Table 7). Among these six missing precursors (Cl, K, Biotin, Heme, Chlorophyll a, R lipoic acids), *Nannochloropsis* is known to take up chloride, potassium and biotin from the growth medium. Hence, exchange reactions were added for these metabolites to import them from the extracellular compartment to the cytosol, where biomass reaction occurs.

Table 7. Missing biomass precursors resulting from the program biomassPrecursorCheck included in COBRA. The function detected six missing precursors. Amongst them, Cl-, K+ and biotin are known to be imported from the medium. The other missing metabolites were fixed by inspecting the pathways that they participate in.

Missing biomass precursors	Approach to produce the missing precursors
Cl^-	Adding exchange reactions
K^+	Adding exchange reactions
Biotin	Adding exchange reactions
Heme	Inspecting porphyrin and chlorophyll biosynthesis pathway
Chlorophyll *a*	Inspecting porphyrin and chlorophyll biosynthesis pathway
R lipoic acids	Adding exchange reactions (required further investigation)

The missing metabolites heme and chlorophyll *a* are produced through the porphyrin and chlorophyll biosynthesis, respectively. Therefore, these pathways were further investigated.

Porphyrin and chlorophyll biosynthesis: Essential genes for the synthesis of chlorophyll a and heme were expected to be present in *Nannochloropsis* [33]. In the initial model, the biosynthesis pathway was found to be incomplete in both cytosol and plastid. Candidate reactions were found and it was checked whether or not their corresponding genes were present in *Nannochloropsis*. The missing reactions were then added from KEGG, and checked against the findings of Vieler and coworkers [33]. As a result, the cytosolic pathway was removed. Eventually, ten reactions were added to the plastid metabolism to complete the pathway.

tRNA is a unique cofactor in this pathway. The requirement of tRNA creates gap in the model thereby blocking the other reactions in the pathway. Therefore, tRNA was removed from the pathway, as its function is merely that of a recycled cofactor. So instead of starting the porphyrin pathway with tRNA glutamate, it was started with glutamate (Figure 32).

The chlorophyll and heme biosynthesis pathway also utilizes Mg and Fe to produce the metabolites magnesium protoporphyrin and heme.Fe, respectively (Figure 32). Mg and Fe were initially not present in the model. Hence, exchange reactions were added for both Mg and Fe.

After this, the model was able to produce chlorophyll a but not heme. The inability to produce heme was caused by the inability of the terpenoid backbone biosynthesis pathway to produce farnesyl diphosphate. Therefore, the the terpenoid pathway was check for the production of the required reactant (Figure 32).

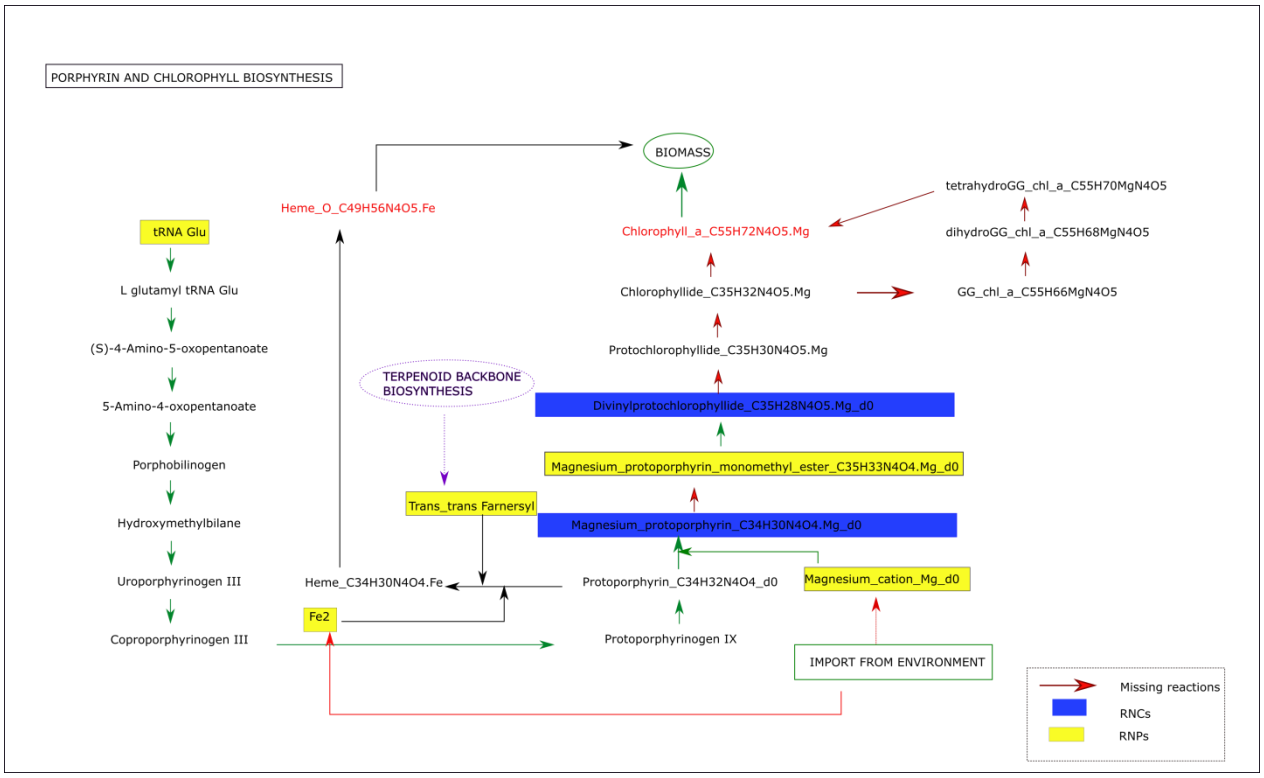


Figure 32. The porphyrin and chlorophyll biosynthesis pathway. The two biomass precursors Chlorophyll a and heme were initially missing because of the presence of five RNPs and two RNCs in this pathway. These gaps were caused by ten missing reactions. Among the gaps, farnesyl was a gap. Since it is produced by another pathway, terpenoid backbone biosynthesis, the later pathway was then investigated.

Terpernoid backbone biosynthesis: Terpenoids is a large class of naturally produced molecules in plants, made from isoprene (C5) units [71]. Terpenoids are associated with both primary and secondary metabolism [71]. Terpenoids can be synthesized via two distinct pathways, the mevalonate (MVA) pathway and the non-mevalonate pathway or the MEP/DOXP pathway [71] [33]. Only the MVA pathway was found in the ancestral eukaryotes, the MEP/DOXP pathway evolved later through the endosymbiotic event where a cyanobacteria was engulfed in another cell, giving rise to glaucophytes, green algae and red algae, the latter being the ancestor of *Nannochloropsis* [33]. In algae like *Nannochloropsis*, all the genes needed to produce isoprenoids can be found in the genome, except for those encoding the enzymes involved in the MVA pathway [33]. The initial model did not support the MVA pathway. Similarly, the MEP/DOXP pathway was incomplete in cytosol and plastid. The initial model did not support the simulation of terpenoid backbone biosynthesis. Hence, the directionality of

certain reactions was reversed. Eight missing reactions in the pathway also generated 4 RNPs and 2 RNCs. Trans-trans farnesyl was among the RNPs and this was also consumed in other reactions within several pathways such as the porphyrin and chlorophyll biosynthesis pathway. Therefore, reactions were added in order to remove the gaps (Figure 33).

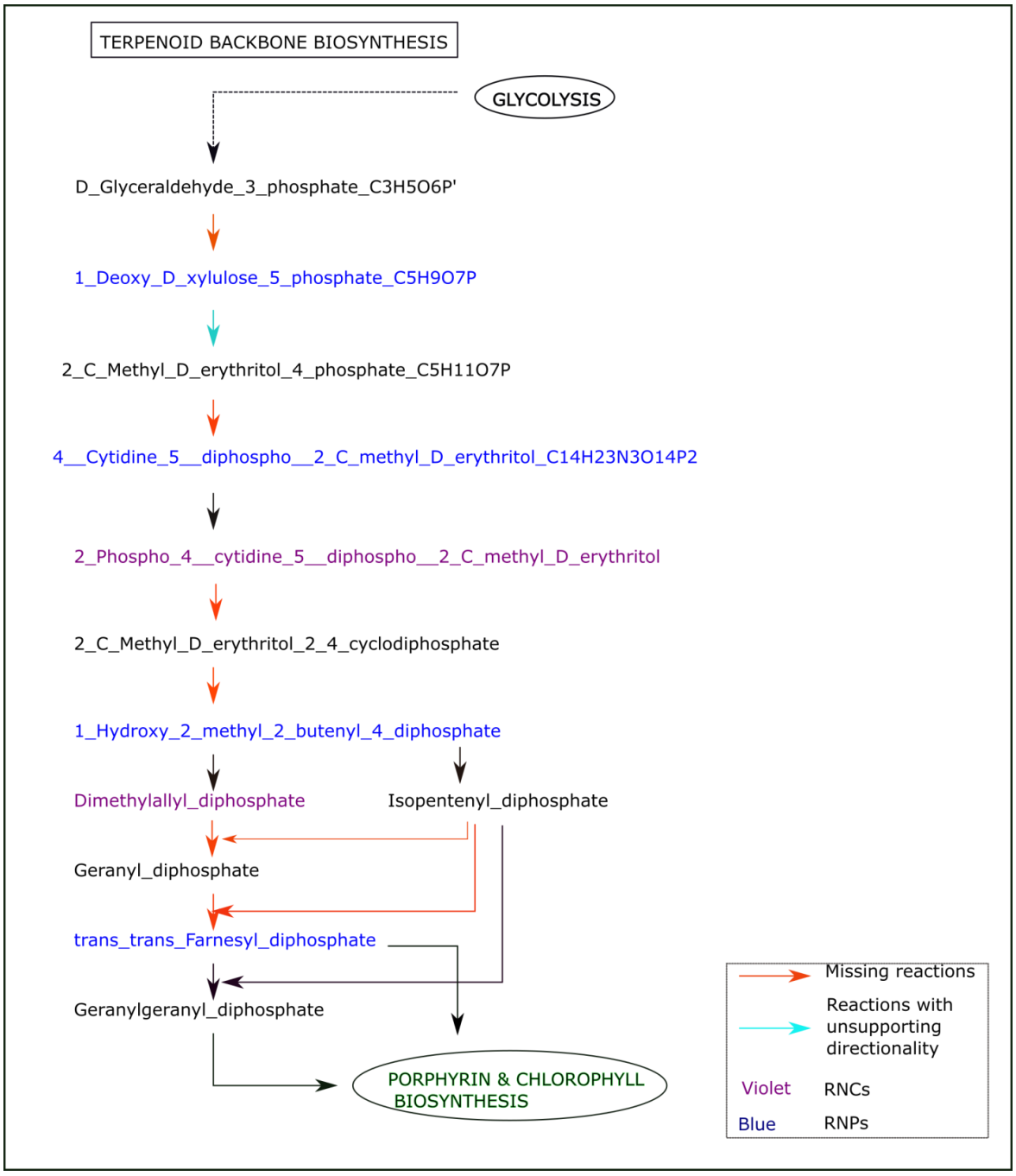


Figure 33. The terpenoid backbone biosynthesis. The pathway contained four RNPs and two RNCs due to the lack of eight reactions and the incorrect directionality of one reaction.

R lipoic acid: was the last missing biomass precursor. It is the product of reaction rxn25981_c0 whose reactant, protein lipoyllysine, is the product of another reaction, rxn25980_c0. However, the reactant of this reaction is again the product of the rxn25981_c0, which creates a cyclic pathway, as shown in Figure 34. As a result, the cyclic pathway produced pseudo gaps of metabolites that are both produced and consumed, but which cannot carry flux. Reactions that include R lipoic acid in databases such as KEGG and Metacyc also include a cyclic pathway. Therefore, in this stage of the modelling, an exchange reaction was added to import R lipoic acid in order to ensure biomass production. Further investigation is required to check the production of R lipoic acid.

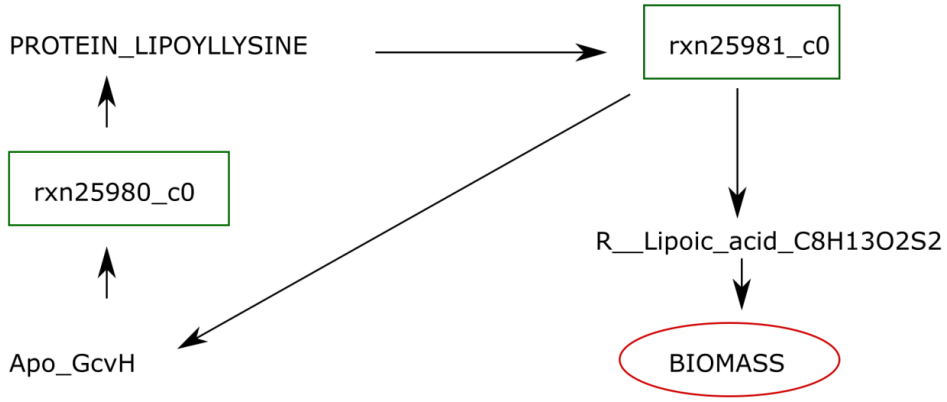


Figure 34. R Lipoic acid production in the initial model. The lipoic acid was involved in a cyclic pathway, and hence, it could not be produced due to the presence of pseudo gaps Apo_GcvH and protein Lipoyllysine

At this stage, no RNPs can be found in the model, and the number of RNCs has been reduced to only 25 metabolites. Some of the RNCs are listed in Table 8 (full list is given in Appendix 2). Most of them resulted from incomplete pathways in incorrect locations. For instance, Palmitoyl-CoA_x0 and Stearoyl-CoA_x0 are two metabolites from the peroxisome that participate in the fatty acid biosynthesis. However, the complete fatty acid biosynthesis is predicted to be located in the plastid in *Nannochloropsis*. Therefore, these two RNCs from the biosynthesis pathway in the peroxisome were not solved. Some RNCs are also involved in pathways that have not been investigated yet. For example L-dehydroascorbate, which is the product in two pigment synthesis reactions described below:

Violaxanthin + L-ascorbate -> H2O + Antheraxanthin + L-dehydroascorbate

L-ascorbate + Antheraxanthin -> H2O+ L-dehydroascorbate + Zeaxanthin

Table 8. The RNCs generated after removal of duplicate reactions. Most of these metabolites are associated with pathways located in incorrect compartments. Some are also involved in pathways that have not yet been investigated. However, none of these currently interfere with the biomass production of the model.

RNCs	Pathways that they interfaced with
Palmitoyl_CoA x0	Fatty acid biosynthesis in peroxisome
Stearoyl_CoA x0	Fatty acid biosynthesis in peroxisome
Dodecanoic_acid_c0	Fatty acid biosynthesis in cytosol
9Z__octadec_9_enoyl_CoA_C39H64N7O17P3S_c0	Fatty acid biosynthesis in cytosol
Palmitoleoyl_ACPs_C16H29OSR_c0	Fatty acid biosynthesis in cytosol
Pyridoxal_phosphate_C8H8NO6P_d0	Alanine, aspartate and glutamate metabolism
1__9Z_octadecenoyl__2_hexadecanoyl_sn_glycero_3_phosphate_C37H69O8P_c0	Glycerophospholipid metabolism
1_2__9Z_octadecenoyl__sn_glycero_3_phosphate_C39H71O8P_d0	Glycerophospholipid metabolism
phosphatidylglycerophosphate__1_18_1_9Z___2_16_1__C40H75O13P2_d0	Glycerophospholipid metabolism
1_2_dipalmitoyl_phosphatidylglycerol_phosphate_C38H73O13P2_c0	Glycerophospholipid metabolism
1_2__9Z_octadecenoyl__sn_glycero_3_phosphate_C39H71O8P_c0	Glycerophospholipid metabolism
Eriodictyol_C15H12O6_c0	Flavonoid biosynthesis
L-dehydroascorbate	Pigment synthesis
Dethiobiotin_C10H17N2O3_c0	Sulfur metabolism
3_Nonaprenyl_4_hydroxybenzoate_C52H77O3_c0	Ubiquinone and other terpenoid-quinone biosynthesis

The presence of these RNCs does currently not affect the biomass production, as no missing biomass precursors were found by the function biomassPrecursorCheck in COBRA. Therefore, these RNCs still remain and will be left for future work.

4.2.2 Biomass precursor check

At the current stage of the modelling, no missing biomass precursor can be found, but still, no biomass production can be obtained when running the function optimizeCbModel in COBRA. The majority of the reactions are still blocked. The biomass precursors were tested again, using a custom made MATLAB program, to see if they are actually produced by the model. In this custom written program, the reactants in biomass reaction were removed and then added one by one, in order to be able to fully control which metabolites that, when added, resulted in zero flux through the biomass reaction. This approach showed that around 70 % (62 precursors) of the total biomass precursors were able to generate a non-zero flux in the biomass reaction (full list given in Appendix 3), thus indicating that the other nineteen biomass precursors might be possible problems for the biomass production. These metabolites are listed in Table 9.

Table 9. The missing biomass precursors found by the custom written MATLAB program for identifying problem metabolites in the biomass reaction. This approach detected 19 missing biomass precursors where the function biomassPrecursorCheck in COBRA were unable to find any. Most of these metabolites are highly connected molecules in the model. But the custom written program also verified that certain gaps are indeed filled. Chlorophyll a was previously detected as a missing metabolite by the biomassPrecursorCheck function. However, after eliminating the gaps in the chlorophyll synthesis pathway, this compound is no longer detected as a missing biomass precursor.

Biomass precursors		FBA result
ATP_C10H13N5O13P3_c0	Zeatin_C10H13N5O_c0	0
NADPH_C21H26N7O17P3_c0	Chlorophyll_a_C55H72N4O5.Mg_c0	0
NADP__C21H25N7O17P3_c0	UDP_6_sulfoquinovose_C15H21N2O19P2S_c0	0
FAD_C27H30N9O15P2_c0	Heme_O_C49H56N4O5.Fe_c0	0
GTP_C10H13N5O14P3_c0	alpha_D_Mannose_C6H12O6_c0	0
FMN_C17H18N4O9P_c0	L_Histidine_C6H9N3O2_c0	0
Thiamin_diphosphate_C12H17N4O7P2S_c0	Ethanolamine_C2H8NO_c0	0
L_Methionine_C5H11NO2S_c0	Hexadecanoic_acid_C16H31O2_c0	0
L_Cysteine_C3H7NO2S_c0	dGTP_C10H13N5O13P3_c0	0
Choline_C5H14NO_c0	L_Histidine_C6H9N3O2_c0	0

Amongst the missing biomass precursors detected by the custom written program, chlorophyll *a*, which was previously detected as a missing reactant when using the function biomassPrecursorCheck in COBRA, was not detected as a missing metabolite anymore. This indeed proved that after inspection and completion of the pathway, the model is able to produce chlorophyll *a*. The biomassPrecursorCheck function in COBRA found missing metabolites by adding new demand reactions for the biomass precursor currently being investigated, and changed the objective function to the currently investigated metabolite, but without altering the biomass reaction. In the cases where the function optimizeCbModel, which was also used by biomassPrecursorCheck, could not produce flux for the new demand reaction of the precursor, the currently investigated metabolite was considered missing. Otherwise, it was considered to be present, and being able to be produced by the model. It is still unclear why the biomassPrecursorCheck function was unable to detect the missing precursors.

However, when changing the objective function the flux distribution also changed. The FBA result showed that only one fourth of the total reactions (265 reactions out of 1025 reactions) were able to carry flux when the objective function was to produce biomass. None of the reaction in the fatty acid biosynthesis in the plastid or the TAG production in the plastid or the peroxisome could carry flux. However, when changing the objective function to the TAG

(instead of only changing the objective function), all reactions in the pathway could carry flux (Figure 15). The total flux carrying reactions in the model was thus increased to 325 reactions.

The overall flux distribution was also changed when changing the objective function to TAG production. Figure 35 shows the fluxes when the biomass reaction is the objective function. Besides the reactions in the fatty acid and TAG biosynthesis pathways, the flux of reactions involving carbon dioxide was also changed (Figure 36). Instead of being imported from the extracellular compartment and used in the plastid, carbon dioxide was produced in the fatty acid biosynthesis and was transported out of the plastid when changing the objective function to TAG production.

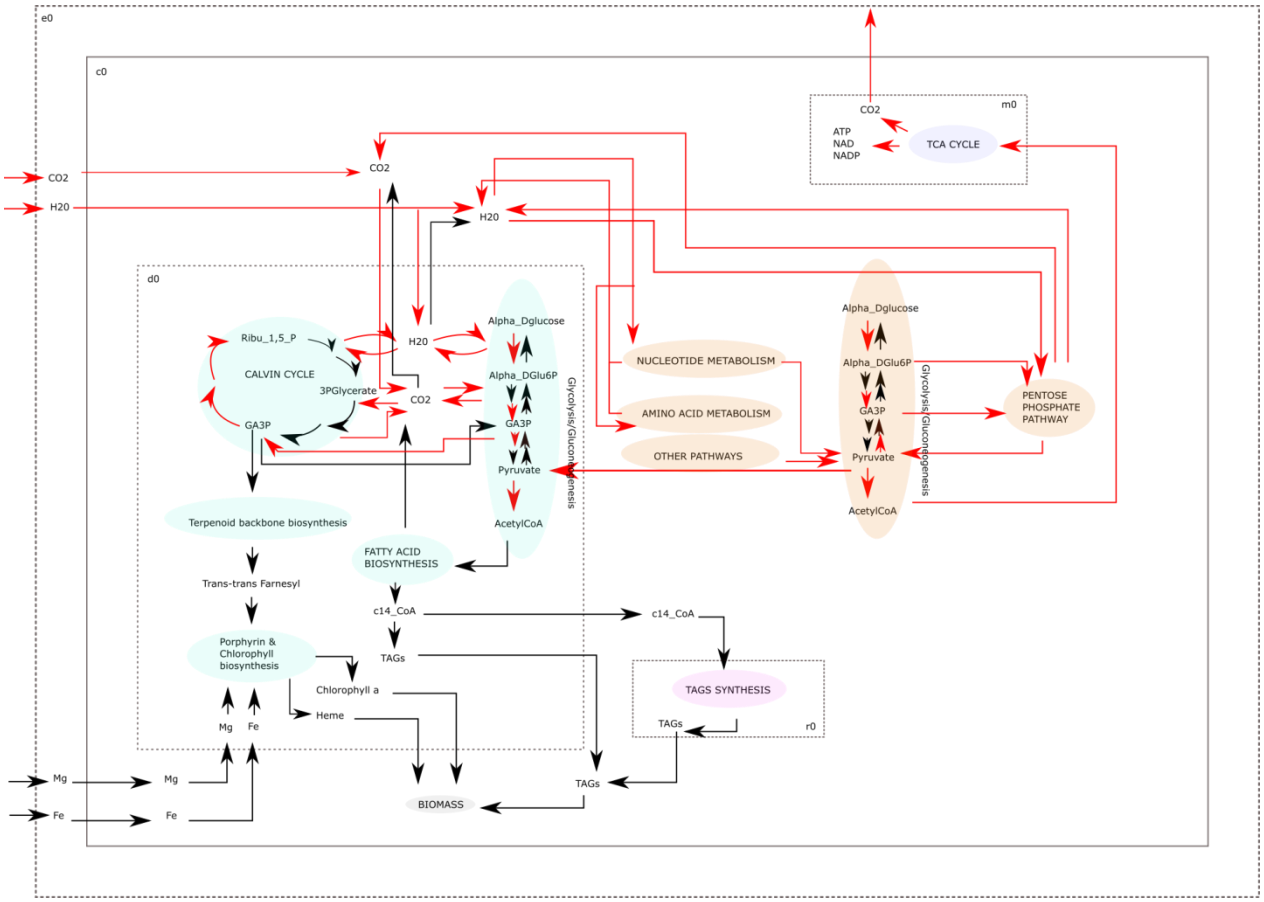


Figure 35. Flux distribution when the biomass reaction is set to be objective function. When biomass is set as objective function, none of the reactions in the fatty acid and the TAG biosynthesis could carry flux. Chlorophyll a and heme can also not be produced.

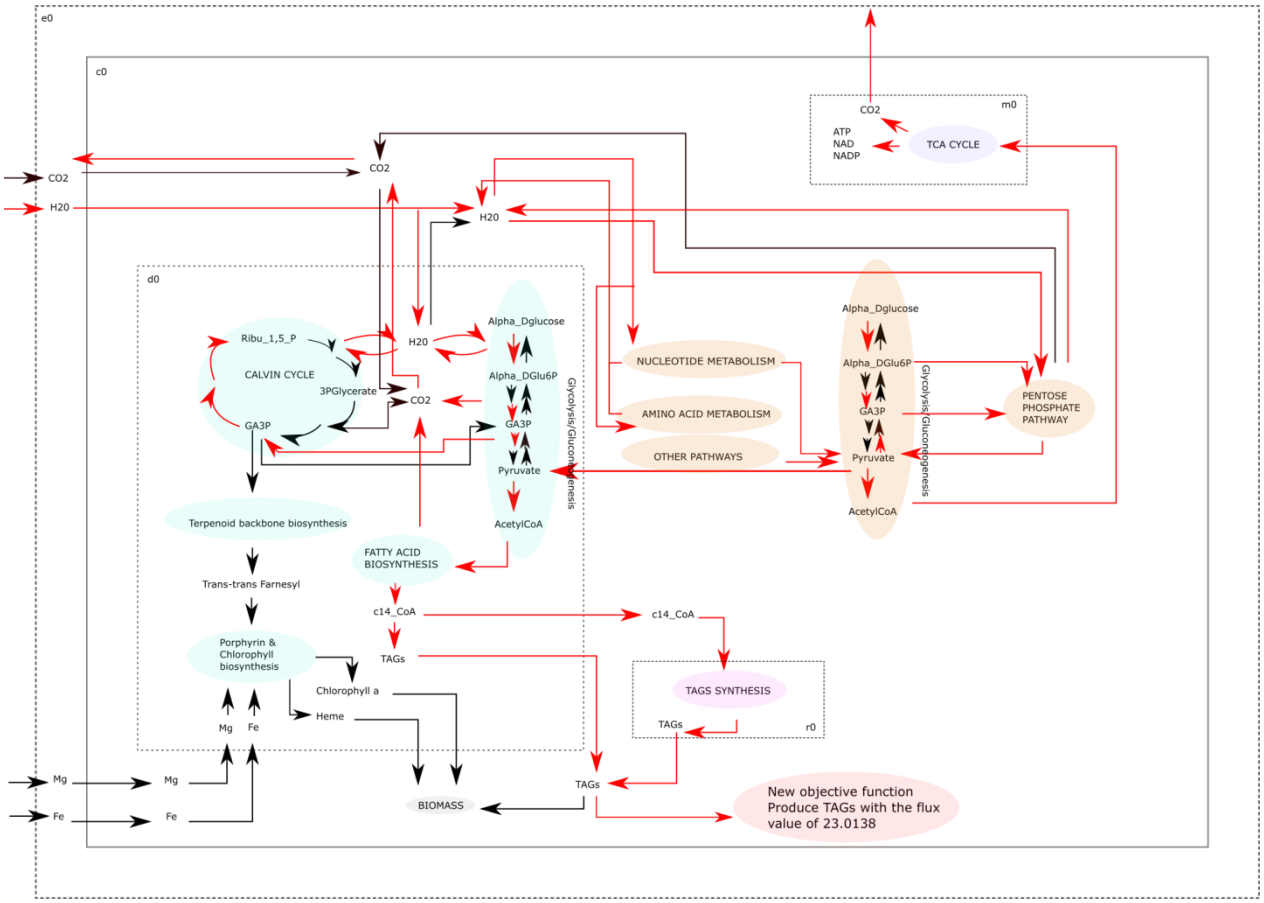


Figure 36. Flux distribution when changing the objective function to TAG production. In this case, all reactions in the fatty acid and TAG biosynthesis pathways could carry flux. TAGs were produced at a rate of 23.0138. Besides the fatty acid and the TAG biosynthesis carrying flux, the flux of the reactions utilizing carbon dioxide were also changed. Instead of CO2 being transported from the extracellular compartment and used in pathways, carbon dioxide was now produced during the production of fatty acid and released to the extra environment.

TAG is one of the 89 biomass precursors. Even though TAG synthesis can produce flux when being set as objective function, it can still not be produced when the objective function is biomass (Figures 35 and 36). It is possible that the reason for the missing flux through the biomass function (although no missing precursors are found by the biomassPrecursorCheck) is that these precursors are not produced when the objective function is biomass production.

Most of the missing precursors have their production reactions blocked, even though all RNCs and RNPs were removed from their production pathways. This is consistent with the work of Thiele, I. and Palsson B.Ø. in 2012. According to their findings, not only dead-end metabolites give rise to gaps, but also unbalanced metabolites. As can be seen in Figure 37, E is an

unbalance metabolite because it is consumed to a larger extent than it is produced. So reaction 3 in Figure 37 cannot carry flux. Hence, metabolite D will not be produced.

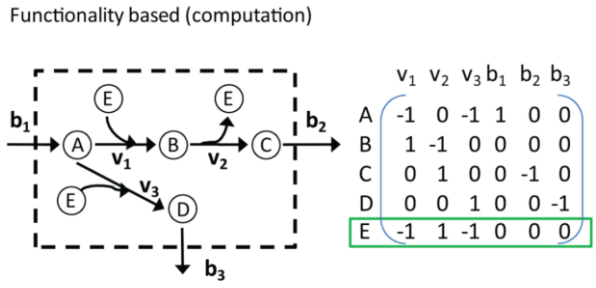


Figure 37. Functionality based gaps. In this example, E is consumed to a higher extent than it is produced. Hence it is an unbalanced metabolite. At steady state, the reactions having fluxes v3 and b3 will be blocked to make sure that the rate at which E is consumed is equal to the rate at which it is produced. The Figure was obtained from Thiele, I. and B.Ø. Palsson [1]

This is consistent with observations made on the model. Energy related molecules such as NADPH, NADH and ATP and some important precursors such as L-glutamate were amongst the most connected nodes in the network, meaning that they participate in many reactions. However, the rate at which they are being produced is much lower than the rate they are consumed at. As a result, they cannot contribute to any significant production of biomass. As an example, ATP in the plastid was only produced in three reactions but was consumed in fifteen reactions. To maintain a steady state where the rate of consumption is equal to the rate of production, only 6 six reactions using ATP can carry flux. Flux variability analysis can be applied to find blocked reactions.

The change in flux distribution observed when changing the objective function to TAG synthesis, and also the presence of functionality gaps, may show that reactions associated with the production of biomass precursors contain unbalanced metabolites. As a result, to obtain a steady state condition, some reactions have to be blocked. At the biomass level, the model can therefore not run fluxes through all the reactions and at the same time produce all the required biomass precursors. The 19 missing precursors found by the alternative program for identifying missing biomass precursors calls for a further investigation, which will be necessary to make the model produce biomass.

4.2.3 Removal of duplicate reactions

The initial model contained many reactions that occurred twice in the model under different IDs. Inspecting the reaction ID could therefore not detect these reactions, and analysis of the stoichiometric matrix was thus necessary. In order to identify overlapping reactions, the stoichiometric matrix was inspected for equal columns. The duplicate reactions are those having the same reactants and products (Figure 38).

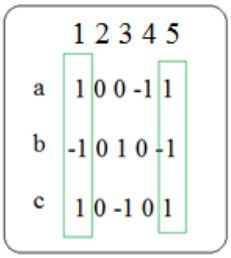


Figure 38. The logics behind the approach used to detect duplicate reactions. The stoichiometric matrix was inspected to find equal columns. In this example reactions 1 and 5 in the matrix above have the same value; which means that these two reactions have the same reactant (b) and the same products (a and c). Hence they are identical.

This inspection resulted in 30 reactions that were found to be included twice in the model (Table 10, full list is given in Appendix 4)

Regardless of the different IDs and enzymes, these reactions appear to have the same equation and genes encoding their enzymes (Table 10). PlantSEED obtained reaction data from many database including KEGG and MetaCyc. These databases were found to contain duplicated reaction [72]. So it may be possible that PlantSEED obtained duplicated reactions in KEGG and MetaCyc when it created the draft reconstruction.

Table 10. Duplicate reactions in the model. Even though the reaction IDs and the enzyme names were different; these reactions have exactly the same reactants and products as well as the genes encoding their enzymes (a full list of duplicate reactions is given in appendix 4)

Reaction1	Reaction2	Reaction equation	Enzyme name for reaction 1	Enzyme name for reaction 2	Genes IDs
rxn00097_c0	rxn27684_c0	1 ATP _c0 + 1 AMP _c0 + 1 H__H_c0 <=> 2 ADP _c0	ATP: AMP phosphotranferase	Adenylate kinase	GI_585105809, GI_585110634, GI_585111217
rxn00423_c0	rxn27108_c0	1 Acetyl_CoA _c0 + 1 L_Serine _c0 <=> 1 CoA _c0 + 1 O_Acetyl_L_serine _c0	acetyl-CoA:L-serine O-acetyltransferase	serine acetyltransferase	GI_585105643

Duplicate reactions would add to the problem of unbalanced metabolites, due to the fact that the metabolites involved in duplicate reactions will be produced or consumed twice as much in the same reaction. All duplicated reactions were removed from the model.

We observed that some reactions retrieved from KBase have identical reactions located in different compartments. In such situation, two scenarios are possible: one of the identical reactions was put in a wrong place making it an un-necessary reaction and needed to be relocated, or it is a duplicate one and needed to be removed. Therefore, these reactions were analyzed. As a result, 16 wrongly located reactions were moved to the location given in KBase. The other 14 duplicate reactions were removed.

This step introduced new metabolites in some compartments. As a result, more root gaps were produced. However, these gaps were not linked to the metabolic pathways that produce biomass precursors in the model. Therefore, it will be left to future work to resolve this.

4.3 Characteristic of the final model

4.3.1 Model overview

The final model characterized most primary metabolic pathways in *Nannochloropsis* in seven compartments including the cytosol, the chloroplast, the mitochondria, the nucleus, the endoplasmic reticulum and the peroxisome. However, I have not analyzed those reactions in the nucleus compartment; hence the final scheme only covers six compartments (Figure 39). As an autotrophic organism, *Nanochloropsis* captures sunlight, water and carbon dioxide to synthesize carbohydrate, i.e.glucose.

Glucose or Glyceraldehyde 3 phosphate (GA3P) was used in the glycolysis to produce pyruvate which is converted to Acetyl-CoA, the initial inputs for TCA cycle. The final outcome of TCA cycle was energy in form of ATP, NADPH, and NADH which in turn used in other biomass precursors biosynthesis pathways.

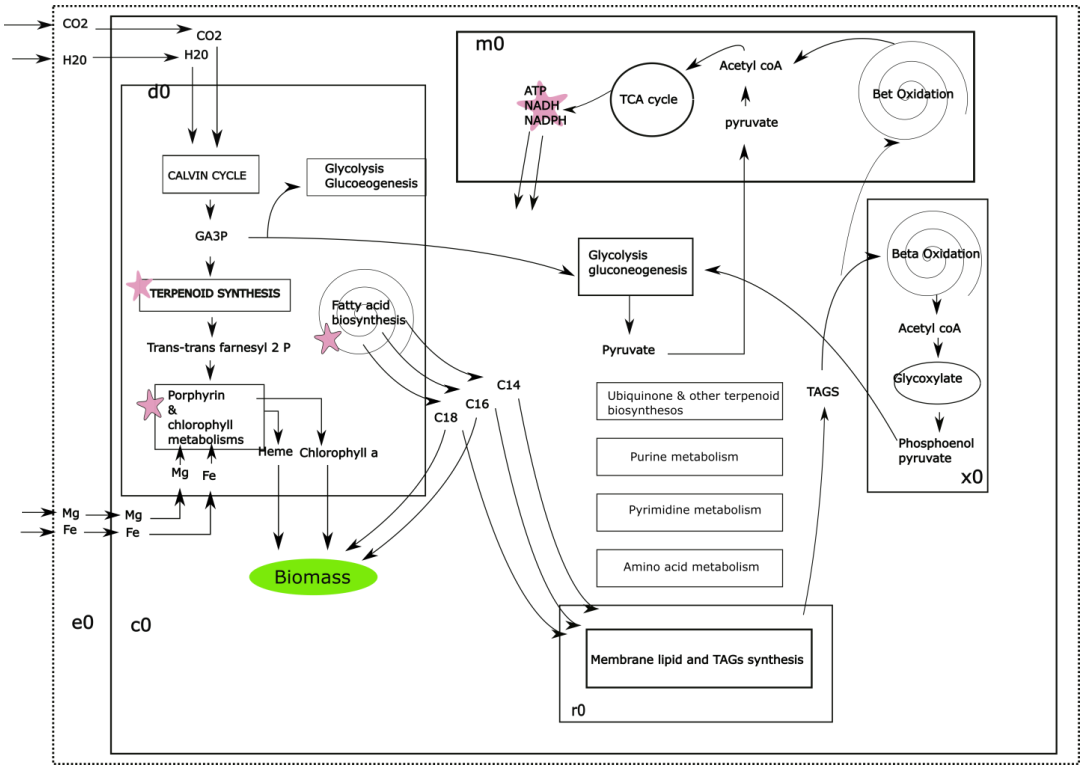


Figure 39. Reconstructed metabolic network of Nannochloropsis. The algae metabolism was characterized in seven compartments. Among them, the reaction in nucleus have not been investigated, hence they are not included in the final scheme. The alga captures sunlight, water and carbon dioxide to synthesize carbohydrate, i.e.glucose. Glucose or Glyceraldehyde 3 phosphate (GA3P) was used in the glycolysis to produce pyruvate which is converted to Acetyl-CoA, the initial inputs for TCA cycle. The final outcome of TCA cycle was energy in form of ATP, NADPH, and NADH which in turn used in other biomass precursors biosynthesis pathways. GA3P- glyceraldehyde 3 phosphate, TAGs- triaccylglyceride; e0_extracellular compartment, c0_the cytosol, d0_the plastid, m0_ the mitochondria, r0_the endoplasmic reticulum, x0_the peroxisome.

The complete scope of the final model includes the Calvin cycle, the TCA cycle, the Glycolysis, the Pentose Phosphate Pathway, Purine and Pyrimidine metabolism, the Terpenoid backbone biosynthesis, Porphyrin and chlorophyll biosynthesis, as well as Fatty acid and TAGs biosynthesis. The other pathways such as Beta oxidation, glycoxylate and amino acid metabolisms are also incompletely included.

The final manually curated genome scale metabolic model of *Nannochloropsis* consists of 1024 metabolites and 987 reactions with 383 genes, which covers 4.8% of the total possible open reading frame (ORFs) in the genome (Table 11). In the reconstruction process, we had to make several assumptions about the presence or absence of the reactions. Although many added reactions had links to enzymes having EC numbers that could be found in *Nanocholoropsis* genome portal [63], I have not located these enzymes's genes due to time limitations. As a result, the number of orphan reactions, or the reactions not being linked to a gene, made up 2/3 of the total reactions in the model. Half of the orphan reactions were added at a later stage, during the manual curation.

Table 11. The characteristic of the final model

Characteristic	***Nannochloropsis* model**
Total possible ORFs	8011
Include genes	383 (4.8% of possible ORFs)
Total reactions	987
Total metabolites	1024
Reactions added to fill gaps by manual curation	364
Non metabolic reactions removed by manual curation	166

Out of 987 reactions in the model, more than 50% of them are in the cytoplasm (Figure 40). This makes the cytosol the largest compartment in the model, which is consistent with the distribution of reactions in previous metabolic models of other photosynthetic organisms, such as for example *Synechocystis* sp. [73] and *Arabidopsis* [14]. However, there may be a bias in defining the cytosolic compartment, since all reactions without location data were added here. This explains the fact that the cytosol accounts for more than 60 % of the metabolites in the model (Figure 40).

Contrary to the metabolic models of other photosynthetic organisms, which highlight mitochondria as the second biggest compartment, in *Nannochloropsis* the second largest hub is the plastid, containing almost 20 percent of the total amount of reactions. Reactions in this compartment participate in carbon fixation through the Calvin cycle, glycolysis, the MVA pathway or terpenoid biosynthesis, Fatty acid biosynthesis, and porphyrin and chlorophyll biosynthesis. These reactions cause 20% of the total metabolites in the model to be present in the plastid. Instead of being the second largest compartment, the mitochondrial compartment in the *Nannochloropsis* model is much smaller than both the plastid and the

cytosol. The compartment is mainly the site for reactions involved in the TCA cycle and metabolism involving various amino acids. *Nannochloropsis* is predicted to carry out beta oxidation in both mitochondria and peroxisome [21]. However, due to the time limitations, I have not analyzed these pathways in the model yet. Future curation needs to be done to complete these compartments. It is also shown that acetyl-CoA generated from beta oxidation in peroxisome is used in the glycoxylate pathway also taken in this compartment [21]. However, at this stage the peroxisome only covers 2% of the total amount of reactions and metabolites in the model, and most of them belong to the metabolism of various amino acids. Both β-oxidation and the glycoxylate pathway are absent in this compartment. Therefore, more reactions need to be added to these pathways in the peroxisome. Similar to the peroxisome, the nucleus is also among the smallest compartments in the model with only 1% of the reactions and 2% of the metabolites located here. The current nuclear reactions also participate in amino acid metabolism. At this stage, the different pathways synthesizing the various amino acids have not yet been verified, so future curation is necessary to analyze and complete these pathways.

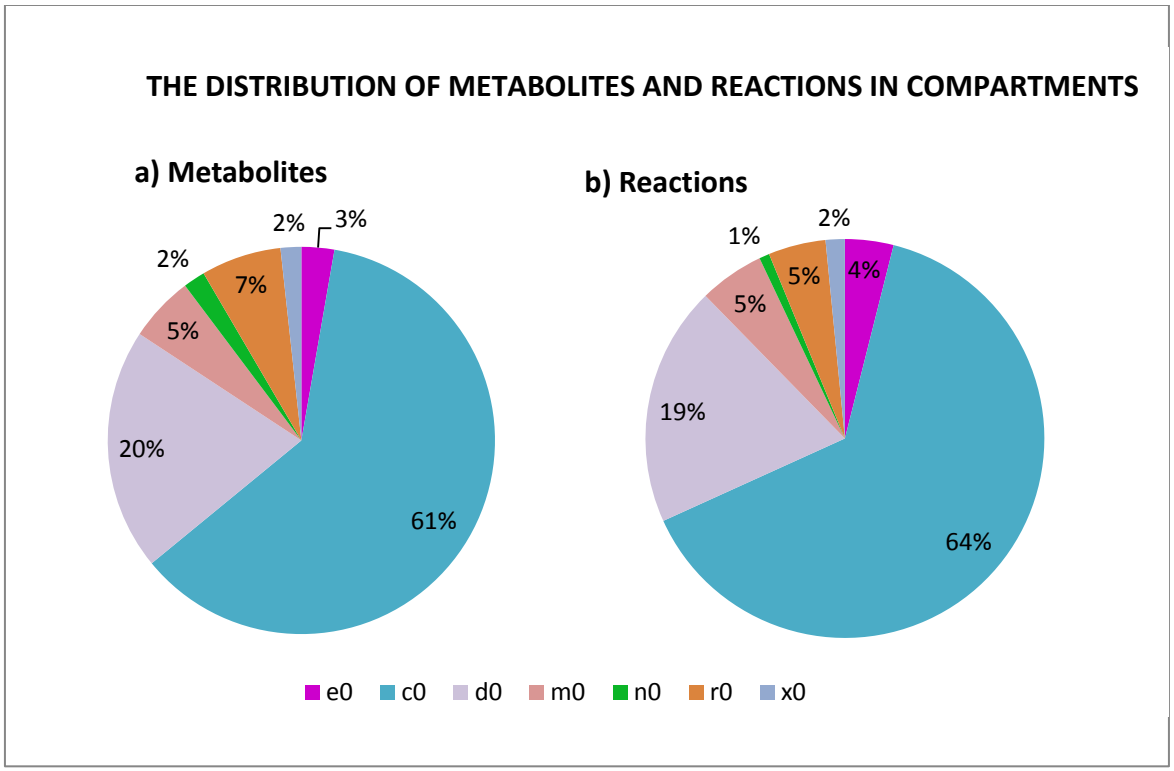


Figure 40. The distribution of reactions and metabolites in the final model. The cytosol is by far the largest compartment in the model, and accounts for more than half of the reactions and metabolites. The cytoplasm is followed in size by the plastid. The number of reactions and metabolites in the mitochondria and endoplasmic reticulum are roughly the same. A minority of the remaining reactions and metabolites are located in the other compartments. In this Figure, e0 denotes the extracellular compartment, c0 denotes the cytosol, d0 denotes the plastid, m0 denotes the mitochondria, n0 denotes the nucleus, r0 denotes the endoplasmic reticulum, and x0 denotes the peroxisome.

The distribution of reactions with respect to metabolic pathways was also analyzed (Figure 41). Similar to preceding metabolic model in other single cell organisms, Henry, C.S. and coworkers [42] and Feist, A.M. and coworkers [74] emphasized the importance of amino acid metabolism and lipid synthesis, and in the current model, a majority of the reactions are involve in these two pathways. One fourth of the total reactions are involved in lipid synthesis, making it the biggest pathway in the model. The large number of reactions involved, indicates a high production of lipid in *Nannochloropsis*, which confirms the finding that these algae can contain up to 60 % lipids [69]. *Nannochloropsis* is predicted to *de novo* synthesize all its amino acids [21], which seems consistent with the overall 18% of the total reactions in the model being involved in biosynthesis of amino acids. The central carbon metabolism involves important carbon fixation processes such as the Calvin-Benson cycle and the C4 cycle; glycolysis; the pentose phosphate pathway, and the TCA cycle, that together make up 9 % of the reactions in the model. These pathways are important metabolic processes that provide energy and crucial precursors for other biomass biosynthesis pathways in the algae. Another large group of reactions were transport reactions, which allow movement of cofactors such as water, carbon dioxide, acetyl-CoA, pyruvate, and energy carriers such as ATP, ADP and NAD, or diffusion of small molecules such as hydrogen, chloride, potassium, iron and magnesium. The remaining reactions are involved in the synthesis of cofactors, vitamins, pigments and several precursors such as terpenoids. Besides, 10 % of the reactions in the model cannot be grouped into distinct pathways. These reactions were so far assigned to an “unknown pathway” and will be inspected later.

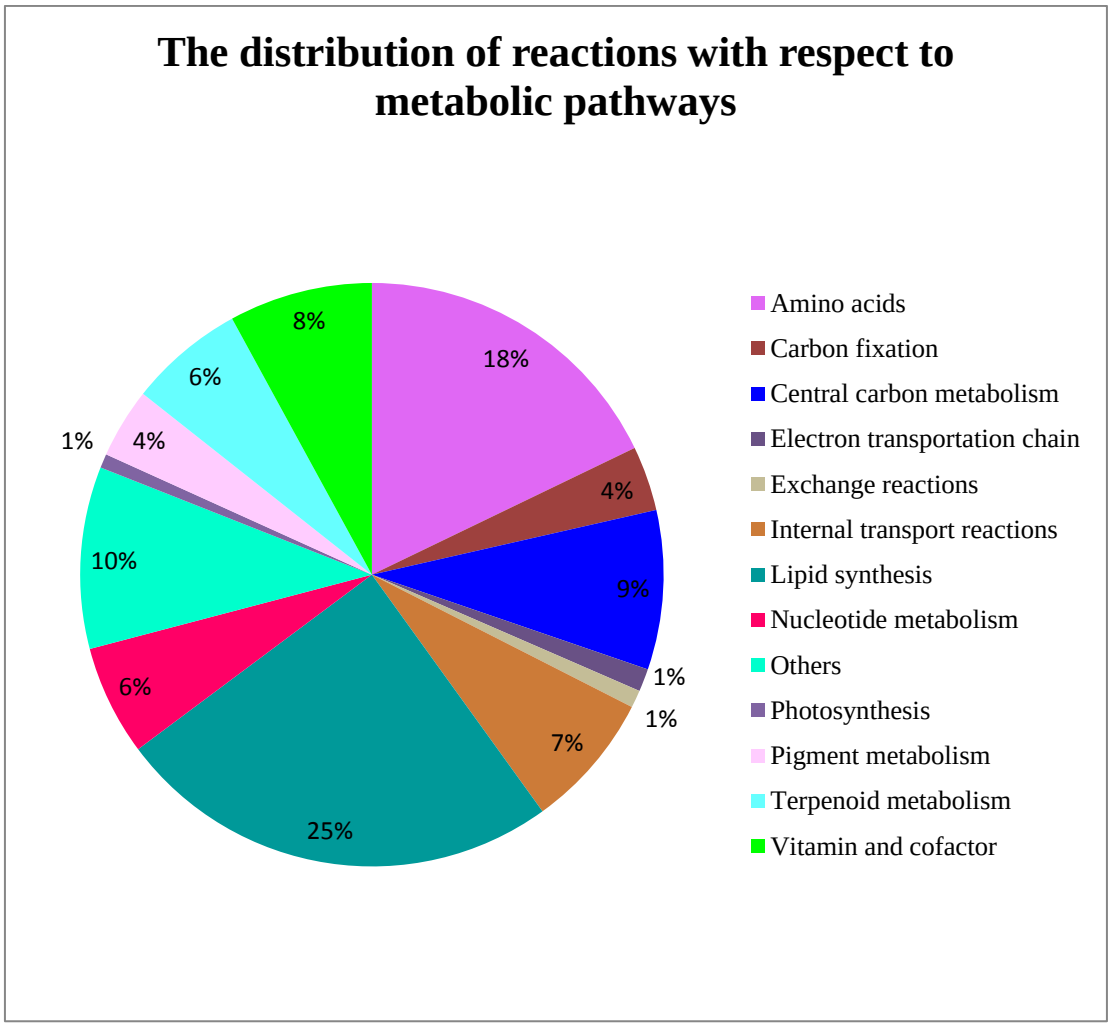


Figure 41. The distribution of reactions with respect to metabolic pathways. The highest numbers of reactions are involved in lipid synthesis and amino acids metabolism, followed by the central carbon metabolism.

4.3.2 Network topology

In the metabolic network, most of the metabolites have very few connections, while very few metabolites are highly connected. These metabolites are referred to as hubs and are crucial to the network [75]. The ranking of most connected metabolites was shown to be the same in most organisms [75]. This means there is a highly conserved network structure within several species [75]. The different that determines the species-specific is the lowly connected or non-hubs metabolites [75].

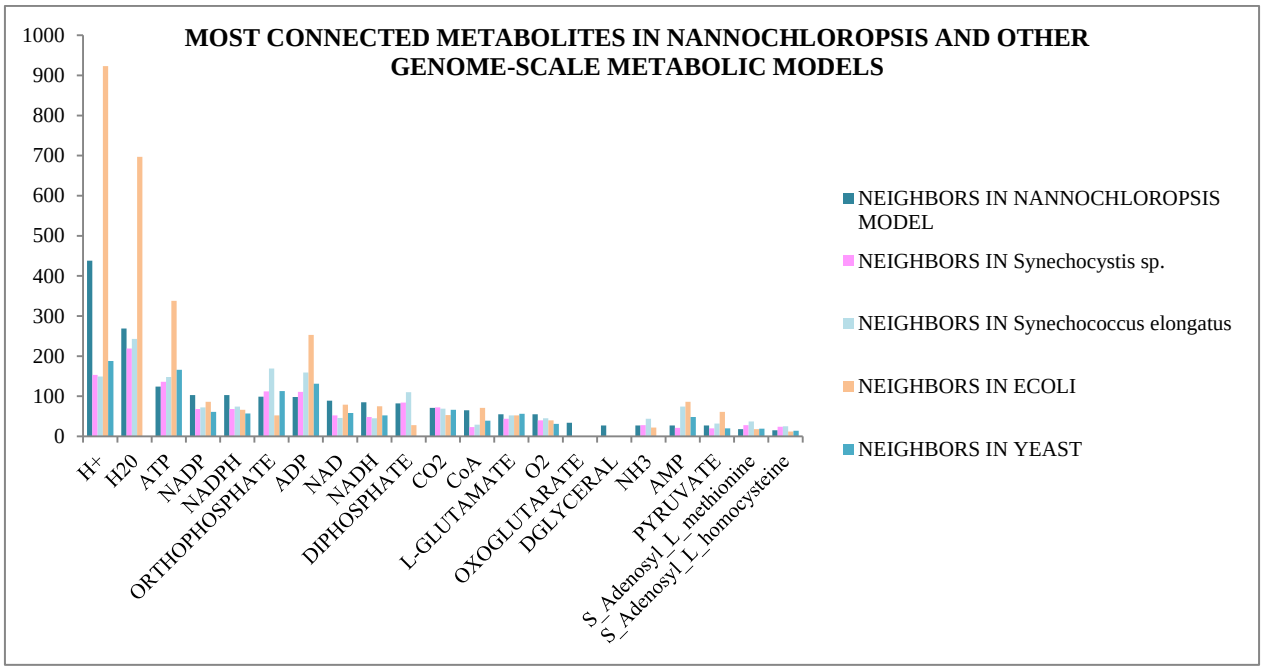


Figure 42. Highly connected metabolites in the final model and in other organisms. The ranking of the most connected nodes in the final Nannochloropsis model is quite similar to the metabolic models of other organisms. The highly connected metabolites are mainly energy carrying molecules and cofactors.

Highly connected molecules in the *Nannochloropsis* model are presented in Figure 42. For the sake of comparison between *Nannochloropsis* and other photosynthetic organisms, the cyanobacterium *Synechococcus elongates* and *Synechocystis sp*. [73] as well as the most precise metabolic models, *E. coli* and yeast were included. The Figure shows that the ranking of highly connected metabolites in *Nannochloropsis* is quite similar to those in other metabolic models (Figure 42). In the metabolic network of *Nannochloropsis*, water is amongst the most connected metabolites. It is not surprising, because water is an important solvent for life due to its role in many types of reactions, such as reduction-oxidation, hydrolysis and condensation. Other highly connected metabolites in *Nannochloropsis* are energy carriers such as ATP, NADH, NAD^+, ADP, orthophosphate, diphosphate and oxygen, followed by amino acids such as L-glutamate and L-glutamine and essential components in porphyrin and chlorophyll metabolism such as S-adenosyl-L-methionine and S-adenosyl-L-homocysteine. In addition, molecules such as ammonia, oxoglutarate, D-glyceraldehyde 3 phosphate, coenzyme A and pyruvate, which are either the substrates or products of many central metabolic pathways such as the TCA cycle,

glycolysis or amino acid metabolism, are also amongst the most connected metabolites in *Nannochloropsis*.

These hubs in the network indicate their potential central roles in the adjustments of fluxes following environmental perturbations and thereby, can be potential candidates for metabolic engineering.

4.4 Reflection on methodologies

Following the guideline for constructing metabolic models [1], I found some challenges when constructing the model of *Nanochloropsis* :

- The function 'biomassPrecursorCheck' in COBRA toolbox failed to find missing biomass precursors in our case, even though several biomass precursors actually were missing. This deviated from what is reported in the guidelines [1]. The reason might be that, when changing the objective function from biomass to the production of each biomass precursor (which the COBRA function does), the flux distribution also changes in order to produce the precursor in question. When the objective function is the biomass, the fluxes are distributed differently than when the objective function is changed to other reactions (in the latter case, one will always achieve the highest possible flux in the reactions producing biomass precursors). As a result, reactions that are involved in producing biomass precursors will not carry flux. This is illustrated in the section 4.4.2. To check the rate of biomass precursor production, I wrote a script with the same algorithm with the 'biomassPrecursorCheck' for checking the production of biomass precursors (shown in Appendix 5). This script detected 19 precursors (out of 89 biomass precursors) being produced at zero flux, which were not found by the original COBRA script. It is still unclear why the COBRA function was unable to detect missing biomass precursors.
- The main source for metabolic information for *Nannochloropsis* is the existing literature, the *Nannochloropsis* genome portal [63] and the KEGG database [28]. There is a lack of experimental biochemical and compartment location data for reactions in this alga, especially in KEGG, which does not consider compartment at all. This is also reported as a problem in previous studies of *Arabidopsis thaliana* [13] and *Chlamydomonas*

reinhardtii [76]. This confirms the observation about the challenges of modelling eukaryote metabolism, due to the lack of metabolic and compartmental data [77].

- Although pathways in KEGG are not organism- specific, the database provides insight into pathways that are present in *Nannochloropsis*. In this work, these reference pathways taken from KEGG were used as a valuable reference when eliminating the gaps in the initial model. Future metabolic modelling projects that want to use data from KEGG for modeling should validate relevant aspects of reactions, such as which reactions that are present in specific organisms, reaction directionality and localization.
- Together with the KEGG database, visualization tools like Cytoscape is a great help when performing gap filling. In this work, Cytoscape was used to visualize pathways and to facilitate inspection of gaps. Based on the schemes of pathways from KEGG, intensive manual layouts are required to visualize the pathway.
- The draft genome scale metabolic reconstruction generated by PlantSEED contained a lot of duplicated reactions and reactions where the compartment identifier of the reaction ID did not correspond to the compartment of the metabolites. The PlantSeed obtained metabolic information from KEGG and Metacyc, which has been reported to contain duplicated reactions [72]. Moreover, the plant-specific nature of the PlantSEED makes it more difficult to apply in modelling alga. Removal of plant specific compartments, such as vacuole, also extends modelling time.
- The flux value has been observed to be difference when using difference solvers, namely Gurobi5 solver and GLPK solver (GNU Linear Programming Kit). As in these solvers, the numbers are rounded differently. Gurobi5 round down those extremely small values, i.e. 10^{-13}, to zeroes, while, GLPK does not. This influenced the investigation of the biomass precursor. As the biomass precursors are missing if their reactions produce zero flux. By contrast, they are present precursors if their reactions produce non-zero flux. A comparison between these two solvers was carried out. Using the same algorithm, Gurobi5 detected 45 missing precursors while GLPK detected only 5 missing precursors. A closer inspection of flux values shown that 40 reactions carry small fluxes in GLPK, have value of zero in Gurobi5 (full list is given in the appendix 6). Although these missing precursors can be detected when inspecting the flux value, yet, during the construction process it is common that the missing precursors often directly generated

without displaying the flux value. Moreover, researcher often uses the same solver when they construct the model. Hence, unless the problem is aware of, the missing precursors will not be detected.

The inconsistency of flux value between different solver was reported by Ebrahim, A. and coworkers in 2015 [78]. They suggested that genome scale metabolic model needs a better standard solver, rather than an exact arithmetic toolbox. However, Chindelevitch, L. and coworkers argued in 2015 [79] that improve standard and arithmetic toolbox are both important. Nevertheless, a better standard solver is apparently essential for analyzing genome scale metabolic models.

4.5 Reflection on findings

- Famili and coworkers stated that the *in silico* model-building procedure at the genome scale is an iterative and ongoing process with continuously growth content [80]. The model refinement stage including the gap filling process can take from months and up to a year to finish [1]. The model is then needed to continuously validation for more accurate simulation. Some models take many years to finish, for example an *Ecoli* model took 13 years to build [81]. This makes it difficult to plan the end-point of the *Nannochloropsis* model reconstruction, which was initiated in this project. During the work done in this thesis, I have completed the gap-filling process. The de-bugging of the initial model has also resolved 70 out of 89 biomass precursors. I estimate that the remaining biomass precursors can be analyzed in two to three months.
- Teusink, B., et al. in 2006 [82] mentioned that the essential parts of metabolic modelling are (i) the development of a quantitative biomass equation, (ii) the analysis and elimination of network gaps, (iii) Determining the stoichiometry of reactions, such as cofactor usage, (iv) determination of stoichiometric coefficients associated with ATP consumption processes not explicitly accounted for in the model. In this thesis, I have eliminated most of network gaps, except 44 new gaps generated from the movement of duplicate reactions from an old compartment to a new one. However, these gaps do not currently interfere with the synthesis of biomass precursors. In order to complete the model, it should be curated further to eliminate these gaps. In this thesis, a biomass equation, covering all the necessary components for algae growth was established with a

suitable stoichiometric coefficient. Stoichiometric coefficients and cofactor usage was also accounted for in most of the reactions in the model. Most of the stoichiometric coefficients were extracted from KEGG. However, according to the finding of Altman, T., et al. in 2013 [72], 1475 (17 % of the total amount of reactions) in the KEGG database are unbalanced. Hence, a future mass and charge balance analysis is required for all reactions in the model.

- At the moment, the model is unable to produce biomass. But when compared to the initial model, which could only produce 9 out of 88 biomass precursors, the model at its current stage can produce 70 out of 89 biomass precursors.
- A positive feature of the constructed model is its ability to produce triacylglycerol (TAGs). TAGs, high energy storage compounds in plants, animal and algae, are of scientific and public interest for their potential to serve as feedstocks and as biofuel [83]. Theoretical prediction for the production of oil in the form of TAGs in *Nannochloropsis* yielded around 40,000 L to 50,000 L /ha/year [84]. The ability to naturally accumulate high content of TAGs in the algae make it a promising biofuel feedstock [19] [21]. The model can serves as a starting point for further validation to achieve more accurate simulation of the TAGs production.

5 Conclusions and future works

This thesis aimed at constructing a genome scale metabolic model of *Nannochloropsis*. The model was constructed from the *Nannochloropsis gaditana* genome, which is the most accurately annotated published genome for this species at the moment. The draft model was generated by PlantSEED and the initial draft was then obtained via KBase. Using the KEGG database, literature and visualization in Cytoscape, I have eliminated all the gaps related to the synthesis of biomass precursors. The subsequent removal of duplicate reactions generated more gaps, but these do not interfere with biomass production. The central metabolic pathways, such as the Calvin cycle, the glycolysis and the TCA cycle have been completed within this project.

The second objective of this thesis was to generate a model that can produce valuable outcomes such as biomass and lipid. Although the reconstruction of biomass is not yet completed, parts of the model are functional, such as the model for lipid production in form of triacylglycerol. The number of missing precursors was reduced significantly from 71 metabolites in the initial draft to 19 metabolites in the final model.

However, due to the limited time to work on the thesis, future work will be required to improve the model:

- The nineteen missing biomass precursors detected later in the work process remains impossible to produce by the current model. Future work is therefore needed to make the model produce these metabolites. As mentioned in section 4.4.2, the presence of missing biomass precursors may be caused by the presence of functionality gaps that block the majority of reactions in the model. Flux variability analysis can be utilized to detect these blocked reactions.
- The oxidative phosphorylation in mitochondria, a metabolic pathway that produce energy in form of ATP molecules and electrons in the form of NADH, is not completed in the model. Further improvement of this pathway may solve the imbalance between the consumption and production of energy related molecules such as ATP and NAD(P)/NAD(P)H. Additional pathways, such as amino acid biosynthesis, vitamin and cofactor biosynthesis also require further revisions as they remain incomplete.
- The presence of cyclic pathways (as the case of R lipoic acid mentioned in section 4.4.2) in the model may produce pseudo-gaps which cannot be detected as RNPs and RNCs,

but are independently presented in the metabolic network. These cyclic pathways will – in agreement with theoretical framework -not carry any flux. Identification and elimination of these cyclic pathways is necessary.

- Gaps are eliminated to facilitate biomass production. Hence, the gap filling process is finished when all biomass precursors can be produced by the biomassPrecursorCheck script. As a result, there still many pathways, for example β-oxidation in mitochondria and peroxisome as well as glycosylate production in peroxisome that need to be investigated and completed.
- A thorough analysis of the mass charge balance for all reactions involved in the biomass production is also required.
- Genes that encode enzymes for all additional and missing reactions should be identified in the genome before the reactions are added.
- The 44 gaps that persist in the model resulted from the relocation of duplicated reactions (section 4.2.3) and need to be removed or linked to the network.

The reconstruction process generates certain limitations. During the gap-filling process, I focused on identifying enzymes that catalyze the gap-fill candidate reactions in the *Nannochloropsis* genome using the genome portal [63] and they are mostly added to the core metabolic pathways, yet, no genes encoding enzymes were assigned. In addition, most metabolic reactions in this genome portal are mainly based on genome annotation. Therefore, it is possible that some reactions will not really occur in the *Nannochloropsis.* Nevertheless, I have included the full list of added and removed reactions with documented subsystems and corresponding enzyme number in the appendix 7, further validation against experimental biochemical data can be done to improve the model.

Eventually, this work presented in this thesis is the first constraint-based, genome-scale metabolic reconstruction of *Nannochloropsis* and provides a comprehensive application of current knowledge concerning the metabolic capabilities of the alga. Several gaps in knowledge, such as experimental biochemical data and location of pathways, were identified for future modeling efforts. The overall aim of the thesis was to reconstruct a genome-scale metabolic network f *Nannochloropsis* and generate a metabolic model from this initial reconstruction. At this stage, the generated model is not able to produce biomass. However, during the periods of

the thesis, the overall aim of the thesis has been achieved. The model constructed in the thesis work can be used as a foundation for further improvement and validation.

6 References

1. Thiele, I. and B.Ø. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nature protocols, 2010. **5**(1): p. 93-121.
2. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends in genetics, 2008. **24**(3): p. 133-141.
3. Fleischmann, R.D., et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science, 1995. **269**(5223): p. 496-512.
4. Metzker, M.L., *Sequencing technologies—the next generation.* Nature reviews genetics, 2010. **11**(1): p. 31-46.
5. KA, W. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).* [cited 2016 26 april]; Available from: https://www.genome.gov/27541954/dna-sequencing-costs/.
6. Wheeler, D.L., et al., *Database resources of the national center for biotechnology information.* Nucleic acids research, 2007. **35**(suppl 1): p. D5-D12.
7. Triana, J., et al., *Generation and evaluation of a genome-scale metabolic network model of Synechococcus elongatus PCC7942.* Metabolites, 2014. **4**(3): p. 680-698.
8. Oberhardt, M.A., B.Ø. Palsson, and J.A. Papin, *Applications of genome-scale metabolic reconstructions.* Molecular systems biology, 2009. **5**(1): p. 320.
9. Raman, K. and N. Chandra, *Flux balance analysis of biological systems: applications and challenges.* Briefings in bioinformatics, 2009. **10**(4): p. 435-449.
10. Auffray, C., et al., *From functional genomics to systems biology: concepts and practices.* Comptes rendus biologies, 2003. **326**(10): p. 879-892.
11. Bordbar, A., et al., *Constraint-based models predict metabolic and associated cellular functions.* Nature Reviews Genetics, 2014. **15**(2): p. 107-120.
12. Boyle, N.R. and J.A. Morgan, *Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii.* BMC systems biology, 2009. **3**(1): p. 1.
13. Poolman, M.G., et al., *A genome-scale metabolic model of Arabidopsis and some of its properties.* Plant physiology, 2009. **151**(3): p. 1570-1581.
14. de Oliveira Dal'Molin, C.G., et al., *AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis.* Plant physiology, 2010. **152**(2): p. 579-589.
15. Amin, S., *Review on biofuel oil and gas production processes from microalgae.* Energy Conversion and Management, 2009. **50**(7): p. 1834-1840.
16. Demirbas, A., *Use of algae as biofuel sources.* Energy conversion and management, 2010. **51**(12): p. 2738-2749.
17. Rodolfi, L., et al., *Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor.* Biotechnology and bioengineering, 2009. **102**(1): p. 100-112.
18. Wang, D., et al., *Nannochloropsis genomes reveal evolution of microalgal oleaginous traits.* PLoS Genet, 2014. **10**(1): p. e1004094.

19. Liu, B., et al., *Triacylglycerol profiling of microalgae Chlamydomonas reinhardtii and Nannochloropsis oceanica.* Bioresource technology, 2013. **146**: p. 310-316.

20. Rocha, J.M., J.E. Garcia, and M.H. Henriques, *Growth aspects of the marine microalga Nannochloropsis gaditana.* Biomolecular engineering, 2003. **20**(4): p. 237-242.

21. Vieler, A., et al., *Genome, functional gene annotation, and nuclear transformation of the heterokont oleaginous alga Nannochloropsis oceanica CCMP1779.* PLoS Genet, 2012. **8**(11): p. e1003064.

22. Jinkerson, R.E., R. Radakovits, and M.C. Posewitz, *Genomic insights from the oleaginous model alga Nannochloropsis gaditana.* Bioengineered, 2013. **4**(1): p. 37-43.

23. Starkenburg, S.R., et al., *A pangenomic analysis of the Nannochloropsis organellar genomes reveals novel genetic variations in key metabolic genes.* BMC genomics, 2014. **15**(1): p. 1.

24. Kremling, A., *Systems biology : mathematical modeling and model analysis*.
ed. Chapman & Hall/CRC mathematical & computational biology series. 2014: Chapman & Hall.

25. Anthony-Cahill, C.K.M.K.E.V.H.D.R.A.S.J., *Biochemistry*. 4th ed. ed. 2013, Toronto, Ont Pearson

26. Schomburg, I., et al., *BRENDA, the enzyme database: updates and major new developments.* Nucleic acids research, 2004. **32**(suppl 1): p. D431-D433.

27. Caspi, R., et al., *The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic acids research, 2008. **36**(suppl 1): p. D623-D631.

28. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.

29. DeBerardinis, R.J. and C.B. Thompson, *Cellular metabolism and disease: what do metabolic outliers teach us?* Cell, 2012. **148**(6): p. 1132-1144.

30. Hu, H. and K. Gao, *Optimization of growth and fatty acid composition of a unicellular marine picoplankton, Nannochloropsis sp., with enriched carbon sources.* Biotechnology letters, 2003. **25**(5): p. 421-425.

31. Hoek, C., D. Mann, and H.M. Jahns, *Algae: an introduction to phycology*. 1995: Cambridge university press.

32. Reyes-Prieto, A., A.P. Weber, and D. Bhattacharya, *The origin and establishment of the plastid in algae and plants.* Annu. Rev. Genet., 2007. **41**: p. 147-168.

33. Radakovits, R., et al., *Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropsis gaditana.* Nature communications, 2012. **3**: p. 686.

34. Lubzens, E., et al., *Potential advantages of frozen algae (Nannochloropsis sp.) for rotifer (Brachionus plicatilis) culture.* Aquaculture, 1995. **133**(3): p. 295-309.

35. Samal, A., *Introduction to Flux Balance Analysis (FBA) and related methods for studying metabolic genotype-phenotype relationships*, in *International Conference on Mathematical and Theoretical Biology*. 2012: Pune, India.

36. Durot, M., P.-Y. Bourguignon, and V. Schachter, *Genome-scale models of bacterial metabolism: reconstruction and applications.* FEMS microbiology reviews, 2009. **33**(1): p. 164-190.

37. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nature Reviews Microbiology, 2009. **7**(2): p. 129-143.

38. Edwards, J.S., R.U. Ibarra, and B.O. Palsson, *In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data.* Nature biotechnology, 2001. **19**(2): p. 125-130.

39. Fell, D.A. and J.R. Small, *Fat synthesis in adipose tissue. An examination of stoichiometric constraints.* Biochemical Journal, 1986. **238**(3): p. 781-786.

40. Majewski, R. and M. Domach, *Simple constrained-optimization view of acetate overflow in E. coli.* Biotechnology and bioengineering, 1990. **35**(7): p. 732-738.

41. Oh, Y.-K., et al., *Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data.* Journal of Biological Chemistry, 2007. **282**(39): p. 28791-28799.

42. Henry, C.S., et al., *iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations.* Genome biology, 2009. **10**(6): p. 1-15.

43. Vinay-Lara, E., et al., *Genome–Scale Reconstruction of Metabolic Networks of Lactobacillus casei ATCC 334 and 12A.* PloS one, 2014. **9**(11): p. e110785.

44. Chatterjee, A. and S. Kundu, *Revisiting the chlorophyll biosynthesis pathway using genome scale metabolic model of Oryza sativa japonica.* Scientific reports, 2015. **5**.

45. Hendry, J.I., et al., *Metabolic model of Synechococcus sp. PCC 7002: Prediction of flux distribution and network modification for enhanced biofuel production.* Bioresource Technology, 2016.

46. Levering, J., et al., *Genome-scale reconstruction of the Streptococcus pyogenes M49 metabolic network reveals growth requirements and indicates potential drug targets.* Journal of biotechnology, 2016.

47. Vongsangnak, W., et al., *Genome-scale metabolic modeling of Mucor circinelloides and comparative analysis with other oleaginous species.* Gene, 2016. **583**(2): p. 121-129.

48. Keseler, I.M., et al., *EcoCyc: a comprehensive database resource for Escherichia coli.* Nucleic acids research, 2005. **33**(suppl 1): p. D334-D337.

49. Cherry, J.M., et al., *SGD: Saccharomyces genome database.* Nucleic acids research, 1998. **26**(1): p. 73-79.

50. Güldener, U., et al., *CYGD: the comprehensive yeast genome database.* Nucleic acids research, 2005. **33**(suppl 1): p. D364-D368.

51. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic acids research, 2007. **35**(suppl 1): p. D26-D31.

52. Peterson, J.D., et al., *The comprehensive microbial resource.* Nucleic acids research, 2001. **29**(1): p. 123-125.

53. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European molecular biology open software suite.* Trends in genetics, 2000. **16**(6): p. 276-277.

54. Markowitz, V.M., et al., *IMG ER: a system for microbial genome annotation expert review and curation.* Bioinformatics, 2009. **25**(17): p. 2271-2278.

55. Ponce-de-León, M., F. Montero, and J. Peretó, *Solving gap metabolites and blocked reactions in genome-scale models: application to the metabolic network of Blattabacterium cuenoti.* BMC systems biology, 2013. **7**(1): p. 114.

56. Schellenberger, J., et al., *BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions.* BMC bioinformatics, 2010. **11**(1): p. 213.

57. Becker, S.A., et al., *Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox.* Nature protocols, 2007. **2**(3): p. 727-738.

58. Guide, M.U.s., *The mathworks.* Inc., Natick, MA, 1998. **5**: p. 333.

59. Gurobi Optimization, I., *Gurobi Optimizer Reference Manual*. 2015.

60. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome research, 2003. **13**(11): p. 2498-2504.

61. Seaver, S. *PlantSEED: A Resource for High-Throughput Comparison, Functional Annotation, and Metabolic Modeling of Plant Genomes*. in *Plant and Animal Genome XXII Conference*. 2014. Plant and Animal Genome.

62. (KBase), D.o.E.S.B.K.; Available from: http://kbase.us.

63. CRIBI, U.o.P.; Available from: http://www.nannochloropsis.org/page/about.

64. Rolfsson, O., B.Ø. Palsson, and I. Thiele, *The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions.* BMC systems biology, 2011. **5**(1): p. 155.

65. Li, J., et al., *Choreography of transcriptomes and lipidomes of Nannochloropsis reveals the mechanisms of oil synthesis in microalgae.* The Plant Cell, 2014. **26**(4): p. 1645-1665.

66. Coleman, R.A. and D.P. Lee, *Enzymes of triacylglycerol synthesis and their regulation.* Progress in lipid research, 2004. **43**(2): p. 134-176.

67. Merchant, S.S., et al., *TAG, You're it! Chlamydomonas as a reference organism for understanding algal triacylglycerol accumulation.* Current opinion in biotechnology, 2012. **23**(3): p. 352-363.

68. Vieler, A., et al., *A lipid droplet protein of Nannochloropsis with functions partially analogous to plant oleosins.* Plant physiology, 2012. **158**(4): p. 1562-1569.

69. Sheehan, J., et al., *A look back at the US Department of Energy's Aquatic Species Program: Biodiesel from algae*. Vol. 328. 1998: National Renewable Energy Laboratory Golden.

70. Moffatt, B.A. and H. Ashihara, *Purine and pyrimidine nucleotide synthesis and metabolism.* The Arabidopsis Book, 2002: p. e0018.

71. Enfissi, E., et al., *Metabolic engineering of the mevalonate and non-mevalonate isopentenyl diphosphate-forming pathways for the production of health-promoting isoprenoids in tomato.* Plant Biotechnology Journal, 2005. **3**(1): p. 17-27.

72. Altman, T., et al., *A systematic comparison of the MetaCyc and KEGG pathway databases.* BMC bioinformatics, 2013. **14**(1): p. 1.

73. Montagud, A., et al., *Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium.* BMC systems biology, 2010. **4**(1): p. 1.

74. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Molecular systems biology, 2007. **3**(1).

75. Silas G. Villas-Boas, J.N., Jorn Smedsgaard, Michael A. E. Hansen, Ute Roessner-Tunali, *Metabolome Analysis: An Introduction*. 2007.

76. de Oliveira Dal'Molin, C.G., et al., *AlgaGEM–a genome-scale metabolic reconstruction of algae based on the Chlamydomonas reinhardtii genome.* BMC genomics, 2011. **12**(4): p. 1.

77. Collakova, E., J.Y. Yen, and R.S. Senger, *Are we ready for genome-scale modeling in plants?* Plant science, 2012. **191**: p. 53-70.

78. Ebrahim, A., et al., *Do genome-scale models need exact solvers or clearer standards?* Molecular systems biology, 2015. **11**(10): p. 831.

79. Chindelevitch, L., et al., *Reply to "Do genome-scale models need exact solvers or clearer standards?".* Molecular systems biology, 2015. **11**(10): p. 830.

80. Famili, I., et al., *Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network.* Proceedings of the National Academy of Sciences, 2003. **100**(23): p. 13134-13139.

81. Reed, J.L. and B.Ø. Palsson, *Thirteen years of building constraint-based in silico models of Escherichia coli.* Journal of Bacteriology, 2003. **185**(9): p. 2692-2699.

82. Teusink, B., et al., *Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model.* Journal of Biological Chemistry, 2006. **281**(52): p. 40041-40048.

83. Hu, Q., et al., *Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances.* The Plant Journal, 2008. **54**(4): p. 621-639.

84. Weyer, K.M., et al., *Theoretical maximum algal oil production.* Bioenergy Research, 2010. **3**(2): p. 204-213.

Appendix 1. A script written based on the gap detection schemes (Figure 20) by Thiele, I. and B.Ø. Palsson [1]

```
% follow the ines thiel et all protocol to find the candidate for gap fill
% in missing biomass precusor. figure 14.
% before run, make sure the latest version of model, missingMets and Deadends are in the
% work space
clear l m n i j k index index2 position consumerxns metsneedtoproducedemand needtoproduce2sink
metsneedtoproduce2demand
clear needtoproduce2sink solution solution1 solution2 solution3 solution4 model_newDemand
model_newDemand2 modelnew modelnew2
l=1;
m=1;
n=1;
for i= 1: length(missingMets)
    position= strmatch(missingMets(i),model.mets); %findposition of the mets
    index= find(model.S(position,:)>0);% find all rxns that produce mets
    for j= 1: length(index)
        metsrxns= find(model.S(:,index(j))~=0); % find all mets in each rxns
        for k= 1: length(metsrxns)
            index2= strmatch(metsrxns(k),Deadends); % check if it is deadend
            if isempty(index2)==1 && model.S(metsrxns(k),index(j))>0; % if it is not deadend
and it is a product
                consumerxns= find(model.S(metsrxns(k),:)<0); % identify the reaction that
consume it
                if isempty(consumerxns)==0 % if there is reaction that consume it
                    % adddemand metsrxns(k)
                    [model_newDemand,addedRxns] =
addDemandReaction(model,model.mets(metsrxns(k)));
                    % maximize for rxn1
                    model_newDemand.c = zeros(length(model_newDemand.c),1);
% CHANGE OBJECTIVE FUNCTION TO NEW DEMAND RXN
    model_newDemand.c(strmatch(addedRxns,model_newDemand.rxns)) = 1;
    solution = optimizeCbModel(model_newDemand);
                    % if flux ~=0, disp metsinrxns(k), find gap fill in
                    if solution.f ~=0
                        metsneedtoproducedemand(l)= metsrxns(k);
                        l= l+1;
                      % else continue with rxns1 at start
                    end
                elseif isempty(consumerxns)== 1 % if there is no reaction that consume it
                    % add sink
                    [modelnew,rxnsInModel] = addSinkReactions(model,model.mets(metsrxns(k)));
                    %minimize for rxns1
                    modelnew.c = zeros(length(modelnew.c),1);
                    modelnew.c(strmatch('sink',modelnew.rxns))=1;
                      solution3 = optimizeCbModel(modelnew,'min');
                     % if flux ~=0, disp metsinrxns(k), find gap fill in
                     if solution3.f ~=0
                         needtoproduce2sink(m)= metsrxns(k);
                         m= m+1;
                    % else continue with rxns1 at start
                     end
                end
            elseif isempty(index2)==0 && model.S(metsrxns(k),index(j))>0; % if it is a deadend
and a product
                % addemand
                [model_newDemand2,addedRxns2] =
addDemandReaction(model,model.mets(metsrxns(k)));
                %maximize for rxn1
                 model_newDemand2.c = zeros(length(model_newDemand2.c),1);
% CHANGE OBJECTIVE FUNCTION TO NEW DEMAND RXN
    model_newDemand2.c(strmatch(addedRxns2,model_newDemand2.rxns)) = 1;
    solution2 = optimizeCbModel(model_newDemand2);
                   % if flux ~=0, disp metsinrxns(k), find gap fill in
                   if solution2.f~=0
                       metsneedtoproduce2demand(m)= metsrxns(k);
                       m=m+1;
```

```
                    % else continue with rxns1 at start
                end
            elseif isempty(index2)==0 && model.S(metsrxns(k),index(j))<0; % if it is a deadend
and not a product
                % add sink
                [modelnew2,rxnsInModel] = addSinkReactions(model,model.mets(metsrxns(k)));
                    %minimize for rxns1
                    modelnew2.c = zeros(length(modelnew2.c),1);
                    modelnew2.c(strmatch('sink',modelnew2.rxns))=1;
                    solution4 = optimizeCbModel(modelnew2,'min');
                     % if flux ~=0, disp metsinrxns(k), find gap fill in
                     if solution4.f ~=0
                         needtoproduce2sink(m)= metsrxns(k);
                         m= m+1;
                    % else continue with rxns1 at start
                     end
            end
        end
    end
end
```

Appendix 2. 25 RNCs after gap filling process

RNCs	Pathways that they interfaced with
Palmitoyl_CoA x0	Fatty acid biosynthesis in peroxisome
Stearoyl_CoA x0	Fatty acid biosynthesis in peroxisome
Dodecanoic_acid_c0	Fatty acid biosynthesis in cytosol
9Z__octadec_9_enoyl_CoA_C39H64N7O17P3S_c0	Fatty acid biosynthesis in cytosol
Palmitoleoyl_ACPs_C16H29OSR_c0	Fatty acid biosynthesis in cytosol
UDP_D_galactose_C15H22N2O17P2_c0	Amino sugar and nucleotide sugar metabolism
GDP_4_dehydro_6_deoxy_D_mannose_C16H21N5O15P2_c0	Amino sugar and nucleotide sugar metabolism
trans_2_Hydroxycinnamate_C9H7O3_c0	Phenylalanine metabolism
Pyridoxal_phosphate_C8H8NO6P_d0	Alanine, aspartate and glutamate metabolism
3-Aminoisobutyric acid; 3-Aminoisobutanoate; 3-Amino-2-methylpropanoate	Pyrimidine degradation, uracil => beta-alanine, thymine => 3-aminoisobutanoate
(R)-5,6-Dihydrothymine;(5R)-Dihydrothymine	Pyrimidine degradation, uracil => beta-alanine, thymine => 3-aminoisobutanoate
1__9Z_octadecenoyl__2_hexadecanoyl_sn_glycero_3_phosphate_C37H69O8P_c0	Glycerophospholipid metabolism
1_2__9Z_octadecenoyl__sn_glycero_3_phosphate_C39H71O8P_d0	Glycerophospholipid metabolism
phosphatidylglycerophosphate__1_18_1_9Z___2_16_1__C40H75O13P2_d0	Glycerophospholipid metabolism
1_2_dipalmitoyl_phosphatidylglycerol_phosphate_C38H73O13P2_c0	Glycerophospholipid metabolism
1_2__9Z_octadecenoyl__sn_glycero_3_phosphate_C39H71O8P_c0	Glycerophospholipid metabolism
Eriodictyol_C15H12O6_c0	Flavonoid biosynthesis
Tetrahydrofolyl__Glu__2__C24H27N8O9_c0	Folate biosynthesis
L-dehydroascorbate	Pigment synthesis
Urea_CH4N2O_c0	Agrinine biosynthesis
Dethiobiotin_C10H17N2O3_c0	Sulfur metabolism
3_Nonaprenyl_4_hydroxybenzoate_C52H77O3_c0	Ubiquinone and other terpenoid-quinone biosynthesis
a_cis_vaccen_2_enoyl__acp__C18H31OSR_c0	MetaCyc Pathway: cis-vaccenate biosynthesis
2__alpha_Hydroxyethyl_thiamine_diphosphate_C14H21N4O8P2S_c0	no pathway
Cytochromes_C_Oxidized_R_d0	no pathway

Appendix 3. All biomass precursors that being produced in the model at the moment

Id	Biomass Precursors	FBA result
1	H2O_H2O_c0	1000
2	NAD__C21H26N7O14P2_c0	-3.411E-13
3	NADH_C21H27N7O14P2_c0	-2.274E-13
4	CoA_C21H32N7O16P3S_c0	1000
5	Pyridoxal_phosphate_C8H8NO6P_c0	1000
6	S_Adenosyl_L_methionine_C15H23N6O5S_c0	45.4545
7	L_Glutamate_C5H8NO4_c0	1000
8	Oxaloacetate_C4H2O5_c0	1000
9	Glycine_C2H5NO2_c0	1000
10	L_Alanine_C3H7NO2_c0	1000
11	L_Lysine_C6H15N2O2_c0	1000
12	L_Aspartate_C4H6NO4_c0	1000
13	L_Arginine_C6H15N4O2_c0	500
14	CTP_C9H13N3O14P3_c0	500
15	L_Glutamine_C5H10N2O3_c0	1000
16	L_Serine_C3H7NO3_c0	1000
17	Ascorbate_C6H7O6_c0	1000
18	UTP_C9H12N2O15P3_c0	666.6667
19	L_Tryptophan_C11H12N2O2_c0	666.6667
20	L_Phenylalanine_C9H11NO2_c0	1000
21	L_Tyrosine_C9H11NO3_c0	1000
22	sn_Glycerol_3_phosphate_C3H7O6P_c0	1000
23	Tetrahydrofolate_C19H21N7O6_c0	1000
24	Choline_C5H14NO_c0	1.0354E-12
25	Cl__Cl_c0	1000
26	Biotin_C10H15N2O3S_c0	1000
27	L_Leucine_C6H13NO2_c0	1000
28	myo_Inositol_C6H12O6_c0	1000
29	5_10_Methylenetetrahydrofolate_C20H21N7O6_c0	1000
30	L_Proline_C5H9NO2_c0	1000
31	S__Malate_C4H4O5_c0	1000
32	L_Asparagine_C4H8N2O3_c0	1000
33	Citrate_C6H5O7_c0	1000
34	L_Valine_C5H11NO2_c0	1000
35	S__Lactate_C3H5O3_c0	1000
36	L_Threonine_C4H9NO3_c0	1000
37	10_Formyltetrahydrofolate_C20H21N7O7_c0	3.1343E-13
38	Potassium_cation_K_c0	1000

39	D_Galacturonate_C6H9O7_c0	1000
40	L_Isoleucine_C6H13NO2_c0	1000
41	cis_Aconitate_C6H3O6_c0	1000
42	5_Methyltetrahydrofolate_C20H23N7O6_c0	1000
43	5_10_Methenyltetrahydrofolate_C20H20N7O6_c0	1000
44	dCTP_C9H13N3O13P3_c0	500
45	dTTP_C10H14N2O14P3_c0	666.6667
46	4_Coumarate_C9H7O3_c0	1000
47	Indole_3_acetate_C10H8NO2_c0	967.9022
48	alpha_D_Galactose_C6H12O6_c0	1000
49	Pantetheine_4__phosphate_C11H21N2O7PS_c0	1000
50	Ferulate_C10H9O4_c0	1000
51	Octadecanoic_acid_C18H35O2_c0	1000
52	beta_Sitosterol_C29H50O_c0	210.1017
53	Campesterol_C28H48O_c0	215.6831
54	beta_Carotene_C40H56_c0	154.1359
55	Xylose_C5H10O5_c0	1000
56	alpha_L_Rhamnose_C6H12O5_c0	1000
57	alpha_Tocopherol_C29H50O2_c0	243.608
58	alpha_L_Arabinose_C5H10O5_c0	1000
59	Folinic_acid_C20H21N7O7_c0	1000
60	Stigmasterol_C29H48O_c0	210.1017
61	Ethylene_C2H4_c0	1000
62	beta_D_Glucopyranuronic_acid_C6H9O7_c0	1000
63	Lutein_C40H56O2_c0	151.567
64	Plastoquinone_9_C53H80O2_c0	125.8687
65	Phytosphingosine_C18H40NO3_c0	125
66	R__Lipoic_acid_C8H13O2S2_c0	1000
67	beta_D_Ribofuranose_C5H10O5_c0	500
68	alpha_D_Glucose_C6H12O6_c0	1000
69	beta_D_Fructose_C6H12O6_c0	1000
70	TAGs_c0	1000

Appendix 4. The full list of duplicated reactions

Id	**rxnsid1**	**rxnsid2overlap**	**Removed or relocate to**
1	rxn00423_c0	rxn27108_c0	Removed
2	rxn00423_c0	rxn27109_c0	Removed
3	rxn00649_c0	rxn19041_c0	Removed
4	rxn00689_c0	rxn27860_c0	Removed
5	rxn00689_c0	rxn27861_c0	Removed
6	rxn00692_c0	rxn19870_c0	Removed
7	rxn00883_c0	rxn28001_c0	Removed
8	rxn00902_c0	rxn27497_c0	Removed
9	rxn01603_c0	rxn19488_c0	Removed
10	rxn19240_c0	rxn27737_c0	Removed
11	rxn19242_c0	rxn27739_c0	Removed
12	rxn19316_c0	rxn27765_c0	Removed
13	rxn20068_c0	rxn23652_c0	Removed
14	rxn20553_c0	rxn20557_c0	Removed
15	rxn27109_c0	rxn00423_c0	Removed
16	rxn27109_c0	rxn27108_c0	Removed
17	rxn00097_c0	rxn27684_c0	d0
18	rxn00148_c0	rxn20509_c0	d0
19	rxn00159_c0	rxn27459_c0	m0
20	rxn00187_c0	rxn19835_c0	d0
21	rxn00187_c0	rxn19836_c0	m0
22	rxn00191_c0	rxn27698_c0	x0
23	rxn00260_c0	rxn27719_c0	d0
24	rxn00260_c0	rxn27720_c0	m0
25	rxn00333_c0	rxn25769_c0	d0
26	rxn00459_c0	rxn17748_c0	d0
27	rxn19835_c0	rxn19836_c0	
28	rxn26238_c0	rxn37803_c0	w0
29	rxn27719_c0	rxn27720_c0	m0
30	rxn37689_c0	rxn37689_m0	m0

Appendix 5. A script written with the same algorithm with 'biomassPrecursorCheck' to analyze the flux of the reaction that produces biomass precursor

```
% the same algorithm with 'biomassPrecursorCheck' which add a new reaction

% to export biomass precursor and change objective function to that new
% reaction. Then run 'optimizeCbModel' to obtain FBA result. If the FBA
result
% is non-zero the biomass precursor is said to be a missing metabolites.
% Otherwise, it is a present metabolite.
clear i index FBA FBAtest
model.c(1) = 0; % set biomass objective function to fail
test = find(model.S(:,1)<0); % extract all reactants (biomass precursors) in
biomass reaction
for i = 1: length(test)
    % add new reaction for the biomass precursors
   model = addReaction(model,'newdemand',model.mets(test(i)),[-1],false);
   % change the objective function to this new reaction
   model.c(end) = 1;
   % run optimizeCbModel to obtain FBA result
   FBA = optimizeCbModel(model);
   FBAtest(i,1) = cellstr(model.metNames(test(i)));
   FBAtest(i,2) = cellstr(num2str(FBA.f));
  end
```

Appendix 6. A comparison of flux value of the biomass precursor reaction obtained from Gurobi5 and GLPK

Id	Biomass precursors	GLPK solver	Gurobi5 solver
1	NAD__C21H26N7O14P2_c0	9.09494701772928e-13	5.68E-13
2	NADH_C21H27N7O14P2_c0	3.48165940522449e-13	0
3	NADPH_C21H26N7O17P3_c0	9.09494701772928e-13	0
4	NADP__C21H25N7O17P3_c0	0	0
5	CoA_C21H32N7O16P3S_c0	-1.31793251048991e-13	0
6	Pyridoxal_phosphate_C8H8NO6P_c0	8.58550232293758e-13	0
7	S_Adenosyl_L_methionine_C15H23N6O5S_c0	-8.2074166116241e-14	0
8	GTP_C10H13N5O14P3_c0	2.59189083201131e-12	0
9	CTP_C9H13N3O14P3_c0	-1.13686837721616e-13	0
10	Thiamin_diphosphate_C12H17N4O7P2S_c0	1.28610671431289e-13	0
11	L_Methionine_C5H11NO2S_c0	-2.51567607082954e-13	0
12	UTP_C9H12N2O15P3_c0	2.29396276097093e-13	0
13	L_Tryptophan_C11H12N2O2_c0	1.0794861808026e-12	-3.4E-13
14	L_Phenylalanine_C9H11NO2_c0	3.50162334454539e-13	0
15	L_Tyrosine_C9H11NO3_c0	-2.63752020315024e-12	0
16	L_Cysteine_C3H7NO2S_c0	2.27373675443232e-13	0
17	Tetrahydrofolate_C19H21N7O6_c0	-1.36424205265939e-12	0
18	Choline_C5H14NO_c0	-7.55226658565375e-12	0
19	dATP_C10H13N5O12P3_c0	1.56702865423686e-13	0
20	L_Histidine_C6H9N3O2_c0	-6.02359570702405e-13	0
21	5_10_Methylenetetrahydrofolate_C20H21N7O6_c0	9.80646746421789e-14	0
22	Ethanolamine_C2H8NO_c0	1.28792027643073e-13	0
23	10_Formyltetrahydrofolate_C20H21N7O7_c0	4.55783932799685e-12	0
24	Zeatin_C10H13N5O_c0	0	0
25	cis_Aconitate_C6H3O6_c0	-5.48158793658402e-13	0
26	5_Methyltetrahydrofolate_C20H23N7O6_c0	-1.50881045932865e-12	0
27	5_10_Methenyltetrahydrofolate_C20H20N7O6_c0	1.02873918912045e-13	0
28	dCTP_C9H13N3O13P3_c0	-9.6540222258397e-13	0
29	dTTP_C10H14N2O14P3_c0	7.52393426199599e-15	0
30	4_Coumarate_C9H7O3_c0	1.66759838748448e-12	0
31	Indole_3_acetate_C10H8NO2_c0	1.19003040932974e-12	0
32	alpha_D_Galactose_C6H12O6_c0	0	0
33	Pantetheine_4__phosphate_C11H21N2O7PS_c0	-1.28755733321031e-12	0
34	Ferulate_C10H9O4_c0	1.35238528994413e-12	0
35	alpha_Tocopherol_C29H50O2_c0	2.34145836029302e-12	0
36	Folinic_acid_C20H21N7O7_c0	-4.74339767424376e-13	0
37	Chlorophyll_a_C55H72N4O5.Mg_c0	4.31031097136586e-14	0

38	Plastoquinone_9_C53H80O2_c0	1.03909362012113e-12	0
39	UDP_6_sulfoquinovose_C15H21N2O19P2S_c0	-9.30173673666485e-13	0
40	Heme_O_C49H56N4O5.Fe_c0	0	0
41	beta_D_Ribofuranose_C5H10O5_c0	-8.47840984343301e-14	0

Appendix 7. The full list of 166 removed reactions

Reaction ID	Reason to remove	Reaction equation
rxn29241_e0	Amonia is not imported	1 NH3_H4N_e0 <=> 1 NH3_H4N_c0
rxn19835_c0	Duplicate reactions	1 ATP_C10H13N5O13P3_c0 + 1 NH3_H4N_c0 + 1 L_Glutamate_C5H8NO4_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 L_Glutamine_C5H10N2O3_c0 + 1 H__H_c0
rxn19836_c0	Duplicated reaction	1 ATP_C10H13N5O13P3_c0 + 1 NH3_H4N_c0 + 1 L_Glutamate_C5H8NO4_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 L_Glutamine_C5H10N2O3_c0 + 1 H__H_c0
rxn20509_c0	Duplicated reaction	1 ATP_C10H13N5O13P3_c0 + 1 Pyruvate_C3H3O3_c0 <=> 1 ADP_C10H13N5O10P2_c0 + 1 Phosphoenolpyruvate_C3H2O6P_c0 + 1 H__H_c0
rxn25769_c0	Duplicated reaction	1 Oxygen_O2_c0 + 1 Glycolate_C2H3O3_c0 -> 1 Hydrogen_peroxide_H2O2_c0 + 1 Glyoxylate_C2HO3_c0
rxn27459_c0	Duplicated reaction	1 NAD__C21H26N7O14P2_c0 + 1 S__Malate_C4H4O5_c0 -> 1 NADH_C21H27N7O14P2_c0 + 1 CO2_CO2_c0 + 1 Pyruvate_C3H3O3_c0
rxn27684_c0	Duplicated reaction	1 ATP_C10H13N5O13P3_c0 + 1 AMP_C10H12N5O7P_c0 + 1 H__H_c0 <=> 2 ADP_C10H13N5O10P2_c0
rxn27698_c0	Duplicated reaction	1 2_Oxoglutarate_C5H4O5_c0 + 1 L_Alanine_C3H7NO2_c0 <=> 1 Pyruvate_C3H3O3_c0 + 1 L_Glutamate_C5H8NO4_c0
rxn27719_c0	Duplicated reaction	1 2_Oxoglutarate_C5H4O5_c0 + 1 L_Aspartate_C4H6NO4_c0 <=> 1 L_Glutamate_C5H8NO4_c0 + 1 Oxaloacetate_C4H2O5_c0
rxn27720_c0	Duplicated reaction	1 2_Oxoglutarate_C5H4O5_c0 + 1 L_Aspartate_C4H6NO4_c0 <=> 1 L_Glutamate_C5H8NO4_c0 + 1 Oxaloacetate_C4H2O5_c0
rxn27860_c0	Duplicated reaction	1 ATP_C10H13N5O13P3_c0 + 1 L_Glutamate_C5H8NO4_c0 + 1 Tetrahydrofolate_C19H21N7O6_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 H__H_c0 + 1 Tetrahydrofolyl__Glu__2__C24H27N8O9_c0
rxn27861_c0	Duplicated reaction	1 ATP_C10H13N5O13P3_c0 + 1 L_Glutamate_C5H8NO4_c0 + 1 Tetrahydrofolate_C19H21N7O6_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 H__H_c0 + 1 Tetrahydrofolyl__Glu__2__C24H27N8O9_c0
rxn37689_m0	Duplicated reaction	1 NADH_C21H27N7O14P2_m0 + 5 H__H_m0 + 1 Ubiquinone_8_C49H74O4_m0 -> 1 NAD__C21H26N7O14P2_m0 + 4 H__H_c0 + 1 ubiquinol_8__C49H76O4_m0
rxn37803_c0	Duplicated reaction	1 sn_Glycerol_3_phosphate_C3H7O6P_c0 + 1 Ubiquinone_8_C49H74O4_c0 -> 1 Glycerone_phosphate_C3H5O6P_c0 + 1 ubiquinol_8__C49H76O4_c0
rxn05330_c0	fatty acid metabolism in cytosol	1 R__3_Hydroxyhexanoyl__acp__C6H11O2RS_c0 <=> 1 H2O_H2O_c0 + 1 trans_Hex_2_enoyl__acp__C6H9ORS_c0
rxn05331_c0	fatty acid metabolism in cytosol	1 R__3_Hydroxydodecanoyl__acp__C12H23O2RS_c0 <=> 1 H2O_H2O_c0 + 1 trans_Dodec_2_enoyl__acp__C12H21ORS_c0
rxn05335_c0	fatty acid metabolism in cytosol	1 3R__3_Hydroxytetradecanoyl__acyl_carrier_protein__C14H27O2RS_c0 <=> 1 H2O_H2O_c0 + 1 trans_Tetradec_2_enoyl__acp__C14H25ORS_c0
rxn05343_c0	fatty acid metabolism in cytosol	1 Octanoyl__acp__C8H15ORS_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 -> 1 CO2_CO2_c0 + 1 3_Oxodecanoyl__acp__C10H17O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0

rxn05345_c0	fatty acid metabolism in cytosol	1 Dodecanoyl__acyl_carrier_protein__C12H23ORS_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 -> 1 CO2_CO2_c0 + 1 3_Oxotetradecanoyl__acp__C14H25O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn05346_c0	fatty acid metabolism in cytosol	1 Butyryl__acp__C4H7ORS_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 -> 1 CO2_CO2_c0 + 1 3_Oxohexanoyl__acp__C6H9O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn05348_c0	fatty acid metabolism in cytosol	1 Decanoyl__acp__C10H19ORS_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 -> 1 CO2_CO2_c0 + 1 3_Oxododecanoyl__acp__C12H21O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn05350_c0	fatty acid metabolism in cytosol	1 Hexanoyl__acp__C6H11ORS_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 -> 1 CO2_CO2_c0 + 1 3_Oxooctanoyl__acp__C8H13O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn05465_c0	fatty acid metabolism in cytosol	1 H__H_c0 + 1 Malonyl_CoA_C24H33N7O19P3S_c0 + 1 Acyl_carrier_protein_HRS_c0 <=> 1 CoA_C21H32N7O16P3S_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0
rxn05736_c0	fatty acid metabolism in cytosol	1 ATP_C10H13N5O13P3_c0 + 1 CoA_C21H32N7O16P3S_c0 + 1 Tetradecanoic_acid_C14H27O2_c0 -> 1 Diphosphate_HO7P2_c0 + 1 AMP_C10H12N5O7P_c0 + 1 H__H_c0 + 1 Tetradecanoyl_CoA_C35H58N7O17P3S_c0
rxn19686_c0	fatty acid metabolism in cytosol	1 3R__3_Hydroxydecanoyl__acyl_carrier_protein__C10H19O2RS_c0 <=> 1 H2O_H2O_c0 + 1 Trans_D2_decenoyl_ACPs_C10H17OSR_c0
rxn25640_c0	fatty acid metabolism in cytosol	1 Hexanoyl__acp__C6H11ORS_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 -> 1 CO2_CO2_c0 + 1 3_Oxooctanoyl__acp__C8H13O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn25642_c0	fatty acid metabolism in cytosol	1 Octanoyl__acp__C8H15ORS_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 -> 1 CO2_CO2_c0 + 1 3_Oxodecanoyl__acp__C10H17O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn25644_c0	fatty acid metabolism in cytosol	1 Decanoyl__acp__C10H19ORS_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 -> 1 CO2_CO2_c0 + 1 3_Oxododecanoyl__acp__C12H21O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn25645_c0	fatty acid metabolism in cytosol	1 Dodecanoyl__acyl_carrier_protein__C12H23ORS_c0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 -> 1 CO2_CO2_c0 + 1 3_Oxotetradecanoyl__acp__C14H25O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn25647_c0	fatty acid metabolism in cytosol	1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 + 1 Myristoyl_ACPs_C14H27OSR_c0 <=> 1 CO2_CO2_c0 + 1 3_Oxohexadecanoyl__acp__C16H29O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn25648_c0	fatty acid metabolism in cytosol	1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 + 1 Myristoyl_ACPs_C14H27OSR_c0 <=> 1 CO2_CO2_c0 + 1 3_Oxohexadecanoyl__acp__C16H29O2RS_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn34722_c0	fatty acid metabolism in cytosol	1 H2O_H2O_c0 + 1 Palmitoyl_ACPs_C16H31OSR_c0 <=> 1 H__H_c0 + 1 Hexadecanoic_acid_C16H31O2_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn34724_c0	fatty acid metabolism in cytosol	1 H2O_H2O_c0 + 1 Octadecanoyl__acyl_carrier_protein__C18H35ORS_c0 -> 1 H__H_c0 + 1 Octadecanoic_acid_C18H35O2_c0 + 1 Acyl_carrier_protein_HRS_c0
rxn37324_c0	fatty acid metabolism in	1 ATP_C10H13N5O13P3_c0 + 1 CoA_C21H32N7O16P3S_c0 + 1 Docosanoic_acid_C22H43O2_c0 -> 1 Diphosphate_HO7P2_c0 + 1

	cytosol	AMP_C10H12N5O7P_c0 + 1 H__H_c0 + 1 Docosanoyl_CoA_C43H74N7O17P3S_c0
rxn05736_r0	fatty acid metabolism in endoplasmic reticulum	1 ATP_C10H13N5O13P3_r0 + 1 CoA_C21H32N7O16P3S_r0 + 1 Tetradecanoic_acid_C14H27O2_r0 -> 1 Diphosphate_HO7P2_r0 + 1 H__H_r0 + 1 AMP_C10H12N5O7P_r0 + 1 Tetradecanoyl_CoA_C35H58N7O17P3S_r0
rxn09449_r0	fatty acid metabolism in endoplasmic reticulum	1 ATP_C10H13N5O13P3_r0 + 1 CoA_C21H32N7O16P3S_r0 + 1 Octadecanoic_acid_C18H35O2_r0 -> 1 Diphosphate_HO7P2_r0 + 1 H__H_r0 + 1 AMP_C10H12N5O7P_r0 + 1 Stearoyl_CoA_C39H66N7O17P3S_r0
rxn37324_r0	fatty acid metabolism in endoplasmic reticulum	1 ATP_C10H13N5O13P3_r0 + 1 CoA_C21H32N7O16P3S_r0 + 1 Docosanoic_acid_C22H43O2_r0 -> 1 Diphosphate_HO7P2_r0 + 1 H__H_r0 + 1 AMP_C10H12N5O7P_r0 + 1 Docosanoyl_CoA_C43H74N7O17P3S_r0
rxn25640_m0	fatty acid metabolism in mitochondria	1 Hexanoyl__acp__C6H11ORS_m0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_m0 -> 1 CO2_CO2_m0 + 1 3_Oxooctanoyl__acp__C8H13O2RS_m0 + 1 Acyl_carrier_protein_HRS_m0
rxn25642_m0	fatty acid metabolism in mitochondria	1 Octanoyl__acp__C8H15ORS_m0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_m0 -> 1 CO2_CO2_m0 + 1 3_Oxodecanoyl__acp__C10H17O2RS_m0 + 1 Acyl_carrier_protein_HRS_m0
rxn25644_m0	fatty acid metabolism in mitochondria	1 Decanoyl__acp__C10H19ORS_m0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_m0 -> 1 CO2_CO2_m0 + 1 3_Oxododecanoyl__acp__C12H21O2RS_m0 + 1 Acyl_carrier_protein_HRS_m0
rxn25645_m0	fatty acid metabolism in mitochondria	1 Dodecanoyl__acyl_carrier_protein__C12H23ORS_m0 + 1 Malonyl__acyl_carrier_protein__C3H3O3RS_m0 -> 1 CO2_CO2_m0 + 1 3_Oxotetradecanoyl__acp__C14H25O2RS_m0 + 1 Acyl_carrier_protein_HRS_m0
rxn25648_m0	fatty acid metabolism in mitochondria	1 Malonyl__acyl_carrier_protein__C3H3O3RS_m0 + 1 Myristoyl_ACPs_C14H27OSR_m0 <=> 1 CO2_CO2_m0 + 1 3_Oxohexadecanoyl__acp__C16H29O2RS_m0 + 1 Acyl_carrier_protein_HRS_m0
rxn25716_m0	fatty acid metabolism in mitochondria	1 H__H_m0 + 1 Malonyl_CoA_C24H33N7O19P3S_m0 + 1 Palmitoyl_ACPs_C16H31OSR_m0 <=> 1 CoA_C21H32N7O16P3S_m0 + 1 CO2_CO2_m0 + 1 3_Oxostearoyl__acp__C18H33O2RS_m0
rxn18900_x0	fatty acid metabolism in peroxisome	1 H2O_H2O_c0 + 1 ATP_C10H13N5O13P3_c0 + 1 Docosanoyl_CoA_C43H74N7O17P3S_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 H__H_c0 + 1 Docosanoyl_CoA_C43H74N7O17P3S_x0
rxn18904_x0	fatty acid metabolism in peroxisome	1 H2O_H2O_c0 + 1 ATP_C10H13N5O13P3_c0 + 1 Palmitoyl_CoA_C37H62N7O17P3S_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 H__H_c0 + 1 Palmitoyl_CoA_C37H62N7O17P3S_x0
rxn18905_x0	fatty acid metabolism in peroxisome	1 H2O_H2O_c0 + 1 ATP_C10H13N5O13P3_c0 + 1 Stearoyl_CoA_C39H66N7O17P3S_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 H__H_c0 + 1 Stearoyl_CoA_C39H66N7O17P3S_x0
rxn18906_x0	fatty acid metabolism in peroxisome	1 H2O_H2O_c0 + 1 ATP_C10H13N5O13P3_c0 + 1 Tetradecanoyl_CoA_C35H58N7O17P3S_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 H__H_c0 + 1 Tetradecanoyl_CoA_C35H58N7O17P3S_x0
rxn15121_m0	Galactose metabolism	1 ATP_C10H13N5O13P3_m0 + 1 alpha_D_Galactose_C6H12O6_m0 -> 1 ADP_C10H13N5O10P2_m0 + 1 H__H_m0 + 1

		alpha_D_Galactose_1_phosphate_C6H11O9P_m0
rxn12282_r0	Glycerophospholipids metabolism	1 ATP_C10H13N5O13P3_r0 + 1 CoA_C21H32N7O16P3S_r0 + 1 9Z__Octadecenoic_acid_C18H33O2_r0 -> 1 Diphosphate_HO7P2_r0 + 1 H__H_r0 + 1 AMP_C10H12N5O7P_r0 + 1 9Z__octadec_9_enoyl_CoA_C39H64N7O17P3S_r0
rxn12282_x0	Glycerophospholipids metabolism	1 ATP_C10H13N5O13P3_x0 + 1 CoA_C21H32N7O16P3S_x0 + 1 9Z__Octadecenoic_acid_C18H33O2_x0 -> 1 Diphosphate_HO7P2_x0 + 1 AMP_C10H12N5O7P_x0 + 1 H__H_x0 + 1 9Z__octadec_9_enoyl_CoA_C39H64N7O17P3S_x0
rxn17469_c0	Glycerophospholipids metabolism	1 CO2_CO2_c0 + 1 Acyl_carrier_protein_HRS_c0 + 1 3_oxo_cis_vaccenoyl_ACPs_C18H31O2SR_c0 -> 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 + 1 Palmitoleoyl_ACPs_C16H29OSR_c0
rxn17469_d0	Glycerophospholipids metabolism	1 CO2_CO2_c0 + 1 Acyl_carrier_protein_HRS_c0 + 1 3_oxo_cis_vaccenoyl_ACPs_C18H31O2SR_c0 -> 1 Malonyl__acyl_carrier_protein__C3H3O3RS_c0 + 1 Palmitoleoyl_ACPs_C16H29OSR_c0
rxn18935_d0	Glycerophospholipids metabolism	1 Beta_3_hydroxybutyryl_ACPs_C4H7O2SR_c0 -> 1 H2O_H2O_c0 + 1 Crotonyl_ACPs_C4H5OSR_c0
rxn18936_c0	Glycerophospholipids metabolism	1 3_Hydroxy_octanoyl_ACPs_C8H15O2SR_c0 <=> 1 H2O_H2O_c0 + 1 trans_Oct_2_enoyl__acp__C8H13ORS_c0
rxn18937_c0	Glycerophospholipids metabolism	1 R_3_Hydroxypalmitoyl_ACPs_C16H31O2SR_c0 <=> 1 H2O_H2O_c0 + 1 trans_Hexadec_2_enoyl__acp__C16H29ORS_c0
rxn19272_d0	Glycerophospholipids metabolism	1 CTP_C9H13N3O14P3_d0 + 1 dipalmitoyl_phosphatidate_C35H67O8P_d0 <=> 1 Diphosphate_HO7P2_d0 + 1 CDP_1_2_dipalmitoylglycerol_C44H79N3O15P2_d0
rxn20553_d0	Glycerophospholipids metabolism	1 sn_Glycerol_3_phosphate_C3H7O6P_d0 + 1 CDP_1_2_dipalmitoylglycerol_C44H79N3O15P2_d0 -> 1 CMP_C9H12N3O8P_d0 + 1 H__H_d0 + 1 1_2_dipalmitoyl_phosphatidylglycerol_phosphate_C38H73O13P2_d0
rxn20557_c0	Glycerophospholipids metabolism	1 sn_Glycerol_3_phosphate_C3H7O6P_c0 + 1 CDP_1_2_dipalmitoylglycerol_C44H79N3O15P2_c0 -> 1 CMP_C9H12N3O8P_c0 + 1 H__H_c0 + 1 1_2_dipalmitoyl_phosphatidylglycerol_phosphate_C38H73O13P2_c0
rxn25662_d0	Glycerophospholipids metabolism	1 R_3_hydroxy_cis_vaccenoyl_ACPs_C18H33O2SR_c0 -> 1 H2O_H2O_c0 + 1 a_cis_vaccen_2_enoyl__acp__C18H31OSR_c0
rxn28072_m0	Glycerophospholipids metabolism	1 sn_Glycerol_3_phosphate_C3H7O6P_m0 + 1 CDP_1_18_1_9Z__2_16_0_glycerol_C46H81N3O15P2_m0 -> 1 CMP_C9H12N3O8P_m0 + 1 H__H_m0 + 1 phosphatidylglycerophosphate__1_18_1_9Z___2_16_1__C40H75O13P2_m0
rxn37217_d0	Glycerophospholipids metabolism	1 Palmitoyl_CoA_C37H62N7O17P3S_c0 + 1 1_Palmitoylglycerol_3_phosphate_C19H37O7P_c0 -> 1 CoA_C21H32N7O16P3S_c0 + 1 dipalmitoyl_phosphatidate_C35H67O8P_c0
rxn37218_d0	Glycerophospholipids metabolism	1 Palmitoyl_CoA_C37H62N7O17P3S_c0 + 1 1_Acyl_sn_glycerol_3_phosphate_C4H8O7PR_c0 -> 1 CoA_C21H32N7O16P3S_c0 + 1 1__9Z_octadecenoyl__2_hexadecanoyl_sn_glycero_3_phosphate_C37H69O8P_c0
rxn00247_n0	incomplete glycolysis in nucleus	1 ATP_C10H13N5O13P3_n0 + 1 Oxaloacetate_C4H2O5_n0 -> 1 CO2_CO2_n0 + 1 ADP_C10H13N5O10P2_n0 + 1 Phosphoenolpyruvate_C3H2O6P_n0
rxn00117_x0	incomplete Purine metabolism in	1 ATP_C10H13N5O13P3_x0 + 1 UDP_C9H12N2O12P2_x0 <=> 1 ADP_C10H13N5O10P2_x0 + 1 UTP_C9H12N2O15P3_x0

	peroxisome	
rxn00237_x0	incomplete Purine metabolism in peroxisome	1 ATP_C10H13N5O13P3_x0 + 1 GDP_C10H13N5O11P2_x0 <=> 1 ADP_C10H13N5O10P2_x0 + 1 GTP_C10H13N5O14P3_x0
rxn00409_x0	incomplete Purine metabolism in peroxisome	1 ATP_C10H13N5O13P3_x0 + 1 CDP_C9H13N3O11P2_x0 <=> 1 ADP_C10H13N5O10P2_x0 + 1 CTP_C9H13N3O14P3_x0
rxn00839_x0	incomplete Purine metabolism in peroxisome	1 ATP_C10H13N5O13P3_x0 + 1 dADP_C10H13N5O9P2_x0 <=> 1 ADP_C10H13N5O10P2_x0 + 1 dATP_C10H13N5O12P3_x0
rxn01353_x0	incomplete Purine metabolism in peroxisome	1 ATP_C10H13N5O13P3_x0 + 1 dGDP_C10H13N5O10P2_x0 <=> 1 ADP_C10H13N5O10P2_x0 + 1 dGTP_C10H13N5O13P3_x0
rxn15010_r0	Maltose metabolism	1 H2O_H2O_r0 + 1 alpha_Maltose_C12H22O11_r0 -> 2 alpha_D_Glucose_C6H12O6_r0
rxn09787_m0	No acetate needed in the mitochondria	1 Acetate_C2H3O2_c0 <=> 1 Acetate_C2H3O2_m0
rxn31198_x0	No citrate needed in the peroxisome	1 Citrate_C6H5O7_c0 <=> 1 Citrate_C6H5O7_x0
rxn09775_m0	No CO2 needed in the mitochondrial	1 CO2_CO2_c0 <=> 1 CO2_CO2_m0
rxn09860_x0	No CO2 needed in the peroxisome	1 CO2_CO2_c0 <=> 1 CO2_CO2_x0
rxn13822_e0	No evidence to find in *Nannochloropsis*	1 5__Deoxyadenosine_C10H13N5O3_c0 -> 1 5__Deoxyadenosine_C10H13N5O3_e0
rxn13286_m0	No Fe needed in the mitochondria	1 Fe2__Fe_c0 <=> 1 Fe2__Fe_m0
rxn28230_e0	No Hydrogen is exported from the metabolism	4 H__H_c0 <=> 4 H__H_e0
rxn29081_e0	No Hydrogen is exported from the metabolism	2 H__H_c0 <=> 2 H__H_e0
rxn31196_x0	No oxygen needed in peroxisome	1 Oxygen_O2_c0 <=> 1 Oxygen_O2_x0
rxn31199_x0	No succinate needed in the peroxisome	1 Succinate_C4H4O4_c0 <=> 1 Succinate_C4H4O4_x0
rxn27709_d0	phenylalanine, tyrosine and tryptophan metabolism	1 L_Glutamine_C5H10N2O3_d0 + 1 Chorismate_C10H8O6_d0 -> 1 Pyruvate_C3H3O3_d0 + 1 L_Glutamate_C5H8NO4_d0 + 1 H__H_d0 + 1 Anthranilate_C7H6NO2_d0
rxn28100_d0	purine in cytosol	1 H2O_H2O_d0 + 1 L_Glutamine_C5H10N2O3_d0 + 1 5_Phospho_alpha_D_ribose_1_diphosphate_C5H9O14P3_d0 -> 1 Diphosphate_HO7P2_d0 + 1 L_Glutamate_C5H8NO4_d0 + 1 H__H_d0 + 1 5_Phosphoribosylamine_C5H11NO7P_d0
rxn05491_e0	Reaction contains incorrect reactants and products	1 H__H_c0 + 1 Adenine_C5H5N5_c0 -> 1 H__H_e0 + 1 Adenine_C5H5N5_e0
rxn09694_e0	Reaction contains incorrect reactants and products	1 H2O_H2O_c0 + 1 ATP_C10H13N5O13P3_c0 <=> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 H__H_e0
rxn10042_c0	Reaction contains incorrect reactants and products	1 H2O_H2O_c0 + 1 ATP_C10H13N5O13P3_c0 + 3 H__H_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 4 H__H_e0

rxn10042_e0	Reaction contains incorrect reactants and products	1 H2O_H2O_c0 + 1 ATP_C10H13N5O13P3_c0 + 3 H__H_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 4 H__H_e0
rxn00001_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 H2O_H2O_w0 + 1 Diphosphate_HO7P2_w0 <=> 1 H__H_w0 + 2 Orthophosphate_HO4P_w0
rxn00189_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 H2O_H2O_w0 + 1 L_Glutamine_C5H10N2O3_w0 <=> 1 NH3_H4N_w0 + 1 L_Glutamate_C5H8NO4_w0
rxn00543_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 NAD__C21H26N7O14P2_w0 + 1 Ethanol_C2H6O_w0 <=> 1 H__H_w0 + 1 NADH_C21H27N7O14P2_w0 + 1 Acetaldehyde_C2H4O_w0
rxn02401_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 ATP_C10H13N5O13P3_w0 + 1 Nicotinate_D_ribonucleoside_C11H13NO6_w0 -> 1 H__H_w0 + 1 ADP_C10H13N5O10P2_w0 + 1 Nicotinate_D_ribonucleotide_C11H12NO9P_w0
rxn09706_g0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 CO2_CO2_c0 <=> 1 CO2_CO2_g0
rxn09874_v0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 H2O_H2O_c0 <=> 1 H2O_H2O_v0
rxn09876_v0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 CO2_CO2_c0 <=> 1 CO2_CO2_v0
rxn21635_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 L_Glutamine_C5H10N2O3_w0 + 1 D_Ribose_5_phosphate_C5H9O8P_w0 + 1 D_Glyceraldehyde_3_phosphate_C3H5O6P_w0 -> 1 H__H_w0 + 3 H2O_H2O_w0 + 1 Orthophosphate_HO4P_w0 + 1 L_Glutamate_C5H8NO4_w0 + 1 Pyridoxal_phosphate_C8H8NO6P_w0
rxn27615_g0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 H2O_H2O_c0 + 1 ATP_C10H13N5O13P3_c0 + 1 Calcium_cation_Ca_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 1 Calcium_cation_Ca_g0 + 1 H__H_c0
rxn27727_v0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 H2O_H2O_v0 + 1 ATP_C10H13N5O13P3_v0 + 3 H__H_v0 -> 1 ADP_C10H13N5O10P2_v0 + 1 Orthophosphate_HO4P_v0 + 4 H__H_c0
rxn31183_v0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 Sulfate_O4S_c0 <=> 1 Sulfate_O4S_v0
rxn31184_v0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 S__Malate_C4H4O5_c0 <=> 1 S__Malate_C4H4O5_v0
rxn31185_v0	Reaction in cell wall (w0), golgi apparatus (g0) and	1 Nitrate_NO3_c0 <=> 1 Nitrate_NO3_v0

	vacuole (v0)	
rxn31186_v0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 Sucrose_C12H22O11_c0 <=> 1 Sucrose_C12H22O11_v0
rxn31187_v0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 Citrate_C6H5O7_c0 <=> 1 Citrate_C6H5O7_v0
rxn32715_v0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 Oxygen_O2_c0 <=> 1 Oxygen_O2_v0
rxn32722_g0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 Oxygen_O2_c0 <=> 1 Oxygen_O2_g0
rxn33113_g0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 Orthophosphate_HO4P_c0 <=> 1 Orthophosphate_HO4P_g0
rxn33306_g0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 H2O_H2O_c0 <=> 1 H2O_H2O_g0
rxn34471_g0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 H__H_g0 <=> 1 H__H_c0
rxn37217_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 Palmitoyl_CoA_C37H62N7O17P3S_w0 + 1 1_Palmitoylglycerol_3_phosphate_C19H37O7P_w0 -> 1 CoA_C21H32N7O16P3S_w0 + 1 dipalmitoyl_phosphatidate_C35H67O8P_w0
rxn37218_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 Palmitoyl_CoA_C37H62N7O17P3S_w0 + 1 1_Acyl_sn_glycerol_3_phosphate_C4H8O7PR_w0 -> 1 CoA_C21H32N7O16P3S_w0 + 1 1__9Z_octadecenoyl__2_hexadecanoyl_sn_glycero_3_phosphate_C37H69O8P_w0
rxn37691_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 NADH_C21H27N7O14P2_c0 + 5 H__H_c0 + 1 Ubiquinone_8_C49H74O4_c0 -> 1 NAD__C21H26N7O14P2_c0 + 4 H__H_w0 + 1 ubiquinol_8__C49H76O4_c0
rxn37803_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 sn_Glycerol_3_phosphate_C3H7O6P_w0 + 1 Ubiquinone_8_C49H74O4_w0 -> 1 Glycerone_phosphate_C3H5O6P_w0 + 1 ubiquinol_8__C49H76O4_w0
rxn15010_w0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 H2O_H2O_w0 + 1 alpha_Maltose_C12H22O11_w0 -> 2 alpha_D_Glucose_C6H12O6_w0
rxn33142_v0	Reaction in cell wall (w0), golgi apparatus (g0) and vacuole (v0)	1 Fe2__Fe_c0 <=> 1 Fe2__Fe_v0
rxn01670_d	secondary	1 H2O_H2O_c0 + 1 Nicotinamide_D_ribonucleotide_C11H14N2O8P_c0

0	metabolism	<=> 1 Orthophosphate_HO4P_c0 + 1 Nicotinamide_beta_riboside_C11H15N2O5_c0
rxn02277_c0	secondary metabolism	1 ATP_C10H13N5O13P3_c0 + 1 CO2_CO2_c0 + 1 7_8_Diaminononanoate_C9H21N2O2_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 Orthophosphate_HO4P_c0 + 3 H__H_c0 + 1 Dethiobiotin_C10H17N2O3_c0
rxn02401_d0	secondary metabolism	1 ATP_C10H13N5O13P3_c0 + 1 Nicotinate_D_ribonucleoside_C11H13NO6_c0 -> 1 ADP_C10H13N5O10P2_c0 + 1 H__H_c0 + 1 Nicotinate_D_ribonucleotide_C11H12NO9P_c0
rxn02401_n0	secondary metabolism	1 ATP_C10H13N5O13P3_n0 + 1 Nicotinate_D_ribonucleoside_C11H13NO6_n0 -> 1 ADP_C10H13N5O10P2_n0 + 1 H__H_n0 + 1 Nicotinate_D_ribonucleotide_C11H12NO9P_n0
rxn02401_x0	secondary metabolism	1 ATP_C10H13N5O13P3_x0 + 1 Nicotinate_D_ribonucleoside_C11H13NO6_x0 -> 1 H__H_x0 + 1 ADP_C10H13N5O10P2_x0 + 1 Nicotinate_D_ribonucleotide_C11H12NO9P_x0
rxn02414_n0	secondary metabolism	1 S_Adenosyl_L_methionine_C15H23N6O5S_n0 + 1 Caffeate_C9H7O4_n0 <=> 1 H__H_n0 + 1 S_Adenosyl_L_homocysteine_C14H20N6O5S_n0 + 1 Ferulate_C10H9O4_n0
rxn02774_c0	secondary metabolism	1 NAD__C21H26N7O14P2_c0 + 1 Precorrin_2_C42H41N4O16_c0 <=> 1 NADH_C21H27N7O14P2_c0 + 2 H__H_c0 + 1 Sirohydrochlorin_C42H38N4O16_c0
rxn03130_c0	secondary metabolism	1 H2O_H2O_c0 + 1 UDP_2_3_bis_3_hydroxytetradecanoyl_glucosamine_C43H75N3O20P2_c0 -> 2 H__H_c0 + 1 UMP_C9H11N2O9P_c0 + 1 Lipid_X_C34H64NO12P_c0
rxn05006_c0	secondary metabolism	1 S_Adenosyl_L_methionine_C15H23N6O5S_c0 + 1 Precorrin_1_C41H38N4O16_c0 <=> 1 S_Adenosyl_L_homocysteine_C14H20N6O5S_c0 + 1 Precorrin_2_C42H41N4O16_c0
rxn05024_d0	secondary metabolism	1 H__H_c0 + 1 2_Succinylbenzoyl_CoA_C32H39N7O20P3S_c0 -> 1 H2O_H2O_c0 + 1 1_4_Dihydroxy_2_naphthoyl_CoA_C32H38N7O19P3S_c0
rxn15348_c0	secondary metabolism	1 H2O_H2O_c0 + 1 Coniferin_C16H22O8_c0 -> 1 beta_D_Glucose_C6H12O6_c0 + 1 Coniferyl_alcohol_C10H12O3_c0
rxn17206_c0	secondary metabolism	1 CO2_CO2_c0 + 1 BCAA_dehydrogenase_2MP_DH_lipoyl_C12H21O2S2R_c0 -> 1 H__H_c0 + 1 3_Methyl_2_oxobutanoic_acid_C5H7O3_c0 + 1 BCAA_dehydrogenase_lipoyl_C8H13OS2R_c0
rxn17206_d0	secondary metabolism	1 CO2_CO2_c0 + 1 BCAA_dehydrogenase_2MP_DH_lipoyl_C12H21O2S2R_c0 -> 1 H__H_c0 + 1 3_Methyl_2_oxobutanoic_acid_C5H7O3_c0 + 1 BCAA_dehydrogenase_lipoyl_C8H13OS2R_c0
rxn17206_n0	secondary metabolism	1 CO2_CO2_n0 + 1 BCAA_dehydrogenase_2MP_DH_lipoyl_C12H21O2S2R_n0 -> 1 H__H_n0 + 1 3_Methyl_2_oxobutanoic_acid_C5H7O3_n0 + 1 BCAA_dehydrogenase_lipoyl_C8H13OS2R_n0
rxn17464_c0	secondary metabolism	1 CoA_C21H32N7O16P3S_c0 + 1 BCAA_dehydrogenase_2MP_DH_lipoyl_C12H21O2S2R_c0 -> 1 2_Methylpropanoyl_CoA_C25H38N7O17P3S_c0 + 1 BCAA_dehydrogenase_DH_lipoyl_C8H15OS2R_c0
rxn21835_c0	secondary	1 Sinapoyl_CoA_C32H42N7O20P3S_c0 + 1

	metabolism	3_hydroxypropyl_glucosinolate_C10H18NO10S2_c0 -> 1 CoA_C21H32N7O16P3S_c0 + 1 3_sinapoyloxypropylglucosinolate_C21H28NO14S2_c0
rxn21837_c0	secondary metabolism	1 S_Benzoate_coenzyme_A_C28H36N7O17P3S_c0 + 1 4_hydroxybutylglucosinolate_C11H20NO10S2_c0 -> 1 CoA_C21H32N7O16P3S_c0 + 1 4_benzoyloxybutylglucosinolate_C18H24NO11S2_c0
rxn21838_c0	secondary metabolism	1 S_Benzoate_coenzyme_A_C28H36N7O17P3S_c0 + 1 2_hydroxy_3_butenylglucosinolate_C11H18NO10S2_c0 -> 1 CoA_C21H32N7O16P3S_c0 + 1 2_benzoyloxy_3_butenylglucosinolate_C18H22NO11S2_c0
rxn21839_c0	secondary metabolism	1 Sinapoyl_CoA_C32H42N7O20P3S_c0 + 1 4_hydroxybutylglucosinolate_C11H20NO10S2_c0 -> 1 CoA_C21H32N7O16P3S_c0 + 1 4_sinapoyloxybutylglucosinolate_C22H30NO14S2_c0
rxn21840_c0	secondary metabolism	1 Sinapoyl_CoA_C32H42N7O20P3S_c0 + 1 2_hydroxy_3_butenylglucosinolate_C11H18NO10S2_c0 -> 1 CoA_C21H32N7O16P3S_c0 + 1 2_sinapoyloxy_3_butenylglucosinolate_C22H28NO14S2_c0
rxn22371_c0	secondary metabolism	1 Z__Phenylacetaldehyde_oxime_C8H9NO_c0 -> 1 N_Benzylformamide_C8H9NO_c0
rxn25978_c0	secondary metabolism	1 CoA_C21H32N7O16P3S_c0 + 1 Pyruvate_dehydrogenase_acetylDHlipoyl_C10H18O2S2R_c0 -> 1 Acetyl_CoA_C23H34N7O17P3S_c0 + 1 Pyruvate_dehydrogenase_dihydrolipoate_C8H16OS2R_c0
rxn26475_c0	secondary metabolism	1 ADP_C10H13N5O10P2_c0 + 2 H__H_c0 + 1 Reduced_NrdH_Proteins_R_c0 -> 1 H2O_H2O_c0 + 1 dADP_C10H13N5O9P2_c0 + 1 Oxidized_NrdH_Proteins_R_c0
rxn26476_c0	secondary metabolism	1 GDP_C10H13N5O11P2_c0 + 2 H__H_c0 + 1 Reduced_NrdH_Proteins_R_c0 -> 1 H2O_H2O_c0 + 1 dGDP_C10H13N5O10P2_c0 + 1 Oxidized_NrdH_Proteins_R_c0
rxn27031_c0	secondary metabolism	2 NADPH_C21H26N7O17P3_c0 + 2 Oxygen_O2_c0 + 2 H__H_c0 + 1 Dihomomethionine_C7H15NO2S_c0 -> 3 H2O_H2O_c0 + 2 NADP__C21H25N7O17P3_c0 + 1 CO2_CO2_c0 + 1 5_Methylthiopentanaldoxime_C6H13NOS_c0
rxn27031_r0	secondary metabolism	2 NADPH_C21H26N7O17P3_r0 + 2 Oxygen_O2_r0 + 2 H__H_r0 + 1 Dihomomethionine_C7H15NO2S_r0 -> 3 H2O_H2O_r0 + 2 NADP__C21H25N7O17P3_r0 + 1 CO2_CO2_r0 + 1 5_Methylthiopentanaldoxime_C6H13NOS_r0
rxn27032_c0	secondary metabolism	2 NADPH_C21H26N7O17P3_c0 + 2 Oxygen_O2_c0 + 2 H__H_c0 + 1 Trihomomethionine_C8H17NO2S_c0 -> 3 H2O_H2O_c0 + 2 NADP__C21H25N7O17P3_c0 + 1 CO2_CO2_c0 + 1 6_Methylthiohexanaldoxime_C7H15NOS_c0
rxn27032_r0	secondary metabolism	2 NADPH_C21H26N7O17P3_r0 + 2 Oxygen_O2_r0 + 2 H__H_r0 + 1 Trihomomethionine_C8H17NO2S_r0 -> 3 H2O_H2O_r0 + 2 NADP__C21H25N7O17P3_r0 + 1 CO2_CO2_r0 + 1 6_Methylthiohexanaldoxime_C7H15NOS_r0
rxn27033_c0	secondary metabolism	2 NADPH_C21H26N7O17P3_c0 + 2 Oxygen_O2_c0 + 2 H__H_c0 + 1 Tetrahomomethionine_C9H19NO2S_c0 -> 3 H2O_H2O_c0 + 2 NADP__C21H25N7O17P3_c0 + 1 CO2_CO2_c0 + 1 7_Methylthioheptanaldoxime_C8H17NOS_c0
rxn27033_r0	secondary metabolism	2 NADPH_C21H26N7O17P3_r0 + 2 Oxygen_O2_r0 + 2 H__H_r0 + 1 Tetrahomomethionine_C9H19NO2S_r0 -> 3 H2O_H2O_r0 + 2 NADP__C21H25N7O17P3_r0 + 1 CO2_CO2_r0 + 1 7_Methylthioheptanaldoxime_C8H17NOS_r0

rxn27349_d0	secondary metabolism	1 H2O_H2O_d0 + 1 ATP_C10H13N5O13P3_d0 + 1 Copper_Cu_d0 <=> 1 ADP_C10H13N5O10P2_d0 + 1 Orthophosphate_HO4P_d0 + 1 Copper_Cu_c0 + 1 H__H_d0
rxn28346_r0	secondary metabolism	1 NADPH_C21H26N7O17P3_r0 + 1 Oxygen_O2_r0 + 1 H__H_r0 + 1 Naringenin_C15H12O5_r0 -> 1 H2O_H2O_r0 + 1 NADP__C21H25N7O17P3_r0 + 1 Eriodictyol_C15H12O6_r0
rxn28615_d0	secondary metabolism	1 Indoleglycerol_phosphate_C11H12NO6P_d0 <=> 1 D_Glyceraldehyde_3_phosphate_C3H5O6P_d0 + 1 Indole_C8H7N_d0
rxn31154_e0	secondary metabolism	1 hn__e0 <=> 1 hn__d0
rxn00577_c0	Sucrose metabolism (not cover in the model)	1 Orthophosphate_HO4P_c0 + 1 Sucrose_C12H22O11_c0 -> 1 D_Fructose_C6H12O6_c0 + 1 D_Glucose_1_phosphate_C6H11O9P_c0
rxn28594_e0	Sucrose metabolism (not cover in the model)	1 Sucrose_C12H22O11_e0 <=> 1 Sucrose_C12H22O11_c0
rxn31200_x0	The product is not needed in the compartment	1 1R_2S__1_Hydroxypropane_1_2_3_tricarboxylate_C6H5O7_c0 <=> 1 1R_2S__1_Hydroxypropane_1_2_3_tricarboxylate_C6H5O7_x0
rxn31201_x0	The product is not needed in the compartment	1 Oxaloacetate_C4H2O5_c0 <=> 1 Oxaloacetate_C4H2O5_x0
rxn31202_x0	The product is not needed in the compartment	1 H__H_c0 <=> 1 H__H_x0
rxn31203_x0	The product is not needed in the compartment	1 CoA_C21H32N7O16P3S_c0 <=> 1 CoA_C21H32N7O16P3S_x0
rxn32972_m0	The product is not needed in the compartment	1 Sulfate_O4S_c0 <=> 1 Sulfate_O4S_m0
rxn32975_x0	The product is not needed in the compartment	1 Sulfate_O4S_c0 <=> 1 Sulfate_O4S_x0
rxn33125_v0	The product is not needed in the compartment	1 Orthophosphate_HO4P_c0 <=> 1 Orthophosphate_HO4P_v0
rxn33423_d0	The product is not needed in the compartment	1 Sucrose_C12H22O11_c0 <=> 1 Sucrose_C12H22O11_d0
rxn34079_m0	The product is not needed in the compartment	1 Nitrate_NO3_c0 <=> 1 Nitrate_NO3_m0
rxn34545_n0	The product is not needed in the compartment	1 Sulfate_O4S_c0 <=> 1 Sulfate_O4S_n0
rxn31205_x0	Vitamin metabolism	1 Hydrogen_peroxide_H2O2_x0 + 1 Ascorbate_C6H7O6_c0 -> 2 H2O_H2O_x0 + 1 Dehydroascorbate_C6H5O6_c0
rxn00029_c0		2 5_Aminolevulinate_C5H9NO3_c0 -> 2 H2O_H2O_c0 + 1 H__H_c0 + 1 Porphobilinogen_C10H13N2O4_c0
rxn00060_c0		1 H2O_H2O_c0 + 4 Porphobilinogen_C10H13N2O4_c0 -> 4 NH3_H4N_c0 + 1 Hydroxymethylbilane_C40H38N4O17_c0
rxn00126_n0		1 H2O_H2O_n0 + 1 ATP_C10H13N5O13P3_n0 + 1 L_Methionine_C5H11NO2S_n0 -> 1 Orthophosphate_HO4P_n0 + 1 H__H_n0 + 1 Diphosphate_HO7P2_n0 + 1

	S_Adenosyl_L_methionine_C15H23N6O5S_n0
rxn00126_r0	1 H2O_H2O_r0 + 1 ATP_C10H13N5O13P3_r0 + 1 L_Methionine_C5H11NO2S_r0 -> 1 Orthophosphate_HO4P_r0 + 1 Diphosphate_HO7P2_r0 + 1 H__H_r0 + 1 S_Adenosyl_L_methionine_C15H23N6O5S_r0
rxn00189_n0	1 H2O_H2O_n0 + 1 L_Glutamine_C5H10N2O3_n0 -> 1 NH3_H4N_n0 + 1 L_Glutamate_C5H8NO4_n0
rxn00830_m0	1 Isopentenyl_diphosphate_C5H10O7P2_m0 <=> 1 Dimethylallyl_diphosphate_C5H10O7P2_m0
rxn00974_c0	1 Citrate_C6H5O7_c0 <=> 1 H2O_H2O_c0 + 1 cis_Aconitate_C6H3O6_c0