Vetle Simensen

Reconstruction of a genome-scale metabolic model for the analysis of the oleaginous phenotype of *Aurantiochytrium* sp. T66

Masteroppgave i MBIOT5
Veileder: Eivind Almaas
Mai 2019

Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for naturvitenskap
Institutt for bioteknologi og matvitenskap

NTNU
Norwegian University of
Science and Technology

Vetle Simensen

Reconstruction of a genome-scale metabolic model for the analysis of the oleaginous phenotype of *Aurantiochytrium* sp. T66

Masteroppgave i MBIOT5
Veileder: Eivind Almaas
Mai 2019

Norges teknisk-naturvitenskapelige universitet
Fakultet for naturvitenskap
Institutt for bioteknologi og matvitenskap

NTNU
Kunnskap for en bedre verden

Acknowledgements

First and foremost, I would like to express my deepest gratitude to everyone in the Network Biology group for making the last couple of years a wonderful experience. The weekly meetings and regular talks have helped tremendously with my work, both by gaining inspiration through the discussion of new approaches, as well as providing valuable feedback on ideas of my own. I would especially like to thank Christian Schulz and Tjasa Kumelj for helping me out with the various obstacles I have stumbled upon during the project. Furthermore, I want to thank professor Eivind Almaas for always believing in me, and for allowing me to take on such a central part in the AurOmega project. The entailing responsibility has allowed me to grow both personally and professionally.

I would also like to thank my fellow students for making these five years in Trondheim the most memorable and fantastic period of my life. This has truly been an exciting journey that I will remember and treasure for the rest of my life. A special thanks to Madeleine and my other study companions for taking part in this experience, it would not have been the same without you.

Finally, I want to thank Edvard and my family for always being there for me. I would not have been able to accomplish this without your unconditional love and support.

Abstract

Thraustochytrids are heterotrophic protists that under certain conditions accumulate large quantities of triacylglycerols (TAGs) rich in ω-3 polyunsaturated fatty acids (PUFAs). The increasing global demand for these PUFA-rich TAGs has consequently made the thraustochytrids being regarded as primary candidates for microbial lipid producing cell factories. However, a systems-level understanding of the metabolic shift from exponential growth to lipid accumulation is in a large extent unclear.

Genome-scale metabolic models (GSMs) allow for a systems-level understanding of the organization and behavior of biochemical networks. The modeling approach integrates all the metabolic capabilities of an organisms within a stoichiometric framework, enabling the *in silico* prediction of the reaction fluxes throughout the metabolic network. The ability to predict cellular phenotypes offers major insights into the properties of metabolic systems, and constraint-based analyses have, for this reason, become one of the most important tools in the systems biological studies of metabolism. GSMs also provide direct correlations between the genotype and biochemical phenotype through boolean gene-reaction associations, allowing for the direct simulation of metabolic engineering strategies.

Using an already published GSM of a closely related strain, a high-quality GSM of the thraustochytrid *Auranthiochytrium* sp. T66 termed iVS1191 was reconstructed. Through iterative refinements and extensive manual curation, the metabolic scope and coverage of the model was significantly improved from that of the template reconstruction. The generated model consisted of 2093 unique metabolic reactions, 1668 metabolites, and 1191 associated genes. Simulated gene essentiality predictions on carbon-limited minimal medium revealed a robust and adaptable metabolic network, able to grow at sub-optimal or optimal growth phenotypes for around 81% of all single-gene knockouts. Using the OptKnock algorithm, multiple double-reaction knockout strategies were proposed for the increased production the key lipid precursors malonyl-CoA and NADPH.

Additionally, the integration of genome-wide transcriptomics data indicate a concurrent down-regulation of specific pathways of the amino acid metabolism at the onset of lipid accumulation, alluding to a conserved transcriptional regulation also observed in other oleaginous microorganisms. The latter analysis also suggest a bimodal regulatory mechanism by which the fatty acid synthase (FAS) complex appears to primarily be regulated on the transcriptional level, while the competing polyketide synthase (PKS) pathway, responsible for the biosynthesis of the ω-3 PUFAs, seems to be regulated on the metabolic level. These results suggest that increasing the intracellular pool of lipid precursor might preferentially benefit the flux through the PKS pathway compared the FAS system, thus increasing the fractional amounts of the ω-3 PUFA in *A.* sp. T66.

Sammendrag

Thraustochytrider er heterotrofe protister som under bestemte betingelser akkumulerer store mengder triglyserider som er svært rike på ω-3-flerumettede fettsyrer. Det økende globale behovet for disse fettsyrene har gjort thraustochytrider til særlig lovende kandidater for mikrobiell produksjon av disse. En forståelse på systemnivå av den metabolske overgangen fra eksponentiell vekst til lipidakkumulering er derimot fortsatt uklar.

Genome-skala metabolske modeller (GSM) muliggjør en forståelse på systemnivå av organiseringen og adferden til biokjemiske nettverk. Denne modelleringsmetoden integrerer samtlige metabolske egenskaper til organismen ved bruk av et såkalt støkiometrisk rammeverk som muliggjør den simultane beregningen av samtlige reaksjonshastigheter i hele det metabolske nettverket. Disse cellulære fenotypene bidrar til stor innsikt i egenskapene til det metabolske systemet, og har av nettopp denne grunnen blitt til et av de viktigste verktøyene innen systembiologisk forskning på metabolske systemer. GSM inkorporerer også direkte forbindelser mellom genotyper og biokjemiske fenotyper ved bruk av boolske gen-reaksjon-assosiasjoner, som muliggjør direkte simulering av knockout-strategier for økt målmetabolittproduksjon.

Ved bruk av en allerede publisert GSM av en nært beslektet stamme, ble en GSM av høy kvalitet rekonstruert for thraustochytriden *Auranthiochytrium* sp. T66. Ved hjelp av iterative modifikasjoner og forbedringer ble det metabolske omfanget og dekningsgraden til modellen utbedret sammenlignet med templatmodellen. Den endelige modellen inneholdt 2093 unike metabolske reaksjoner, 1668 metabolitter, og 1191 tilhørende gener. Simulerte genessensialitetsanalyser på minimalt karbon-medium antydet at det metabolske nettverket er både robust og tilpasningsdyktig, ved å oppnå optimal eller sub-optimal vekst for rundt 81% av alle gen-knockouter. Flere doble reaksjonsmutanter ble identifisert ved bruk av OptKnock-algoritmen som resulterte i økte produksjonsrater av de essensielle lipidforløperne malonyl-CoA og NADPH.

Ved inkorporering av transkriptomdata ble det også oppdaget en sammenfallende nedregulering av spesifikke reaksjonsspor i aminosyremetabolismen ved nitrogen-begrensing, noe som tyder på en konservert reguleringsmekanisme som også er observert i andre oljeholdige mikroorganismer. Den sistnevnte analysen antyder også en bimodal reguleringsmekanisme der fettsyrasyntase-komplekset (FAS) virker å være primært regulert på transkriptomnivå, mens det konkurrerende enzymkomplekset ansvarlig for produksjnoen av de ω-3-flerumettede fettsyrene, polyketid-syntase (PKS), tilsynelatende er regulert på metabolittnivå. Dette kan tyde på at en økning av den intracellulære mengden av lipidforløperne vil kunne øke aktiviteten til PKS, og i mindre grad påvirke aktiviteten til FAS, som følgelig vil kunne forbedre den fraksjonelle sammensetningen av de ønskede ω-3-fettsyrene i *A.* sp. T66.

Table of Contents

List of Tables

List of Figures

Abbreviations

ACC1	=	Acetyl-CoA carboxylase
ACL	=	ATP-citrate lyase
ACP	=	Acyl-carrier protein
BLAST	=	Basic Local Alignment Search Tool
CB	=	Convex Basis
CoA	=	Coenzyme A
COBRA	=	Constraint-Based Reconstruction and Analysis
DHA	=	Docosahexaenoic acid
EC	=	Enzyme commision
EGC	=	Energy-generating cycle
FAS	=	Fatty acid synthase
FBA	=	Flux balance analysis
FVA	=	Flux variability analysis
GAM	=	Growth associated maintenance
GSM	=	Genome-scale metabolic model
G6PD	=	Glucose-6-phosphate 1-dehydrogenase
HMM	=	Hidden markov model
KEGG	=	Kyoto Encyclopedia of Genes and Genomes
KO	=	KEGG Orthology
Mb	=	Mega base pairs
MC	=	Mitochondrial carrier
ME	=	Malic enzyme
MILP	=	Mixed-integer linear programming
MOMA	=	Minimization of metabolic adjustment
NGAM	=	Non-growth associated maintenance
ODE	=	Ordinary differential equation
ORF	=	Open reading frame
PGD	=	6-phosphogluconate dehydrogenase
PGK	=	Phosphoglycerate kinase
PKS	=	Polyketide synthase
PPP	=	Pentose phosphate pathway
PUFA	=	Polyunsaturated fatty acid
RAVEN	=	Reconstruction, analysis, and visualization of metabolic networks
RPIA	=	Ribose-5-phosphate isomerase
S. limacinum	=	*Schizochytrium limacinum*
TAG	=	Triacylglycerol
TCA	=	Tricarboxylic acid
TCDB	=	Transporter classification database

Chapter 1

Introduction

For several decades, accumulated evidence has been gathered for the health benefits of human consumption of marine sea foods [1]. The high content of ω-3 polyunsaturated fatty acids (PUFAs), such as docosahexaenoic acid (DHA), is regarded as one of the major contributors to this effect [1, 2]. These fatty acids are necessary for the proper development of the fetal brain and retina [3, 4], and have shown to possess antiinflammatory properties associated with reduced risks of cardiovascular diseases [5, 6]. Although rich in ω-3 fatty acids, oily fish such as salmon, mackerel and herring do not synthesize these *de novo*, but rather acquire them from feeding on lipid-rich plankton and microalgae [1]. The rising global aquaculture industry, particularly the marine pisciculture industry, has consequently led to an escalating need for novel sources of these fatty acids [7], whose current source is predominately that of fish oil [8]. Owing to their trophic level in the food chain, these PUFA-containing lipids contain elevated levels of toxic contaminants such as polychlorinated biphenyls and heavy metals [9, 10]. A need for alternative sources of high-quality ω-3 PUFAs is therefore of utmost importance.

Cultivation and strain engineering of lipid-producing microorganisms is regarded as a promising strategy to cater for this demand. Despite the fact that a large variety of phylogenetically diverse microorganisms accumulate substantial amounts of lipids, only a limited number of oleaginous species produce lipids rich in ω-3 PUFAs [11]. One of these ω-3 PUFA-producing organisms are a group of marine heterokonts called thraustochytrids. The thraustochytrids are a taxinomically ambiguous group of heterotrophic unicellular protists that accrue large quantities of ω-3-rich triacylglycerols (TAGs) [12]. These TAGs are stored as intracellular lipid droplets that may constitute up to 80% of the total cell mass, and have DHA contents of up to 80%, albeit not simultaneously [13, 14].

In most PUFA-producing microorganisms, the fatty acids are produced by elongating and desaturating the products of the fatty acid synthase (FAS) enzymatic machinery [15]. However, the thraustochytrids predominately synthesize their PUFAs using a competing polyketide synthase (PKS) system. Here, the acyl intermediates may retain their unsaturated bonds during the biosynthetic process, thus lowering the molar requirement for reducing power compared to the traditional elongation/desaturation scheme [16]. The

capability to accumulate large amounts of ω-3 PUFAs, as well as the decreased need for reducing power, has made thraustochytrids particularly promising candidates for effective lipid-producing cell factories [16].

To elucidate the biological mechanisms underpinning the fatty acid biosynthesis of thraustochytrids, knowledge of the global metabolic organization and functionality is vital. Metabolic modeling has become a pivotal methodology to simulate and predict metabolic phenotypes *in silico*, thereby generating hypotheses and guide experimental efforts [17]. Whereas some modeling approaches require extensive parametrization, constraint-based modeling using genome-scale metabolic reconstructions, in its most fundamental form, merely calls for an annotated genome sequence [18]. Applying simple assumption of mass balance, a steady state, and an evolutionary motivated objective function to be optimized, this approach allows for rapid and efficient predictions of metabolic fluxes throughout the metabolic network [19].

The principal aim of this thesis was to generate a high-quality genome-scale metabolic model (GSM) of the thraustochytrid *Aurantiochytrium* sp. T66.

This was performed by using an already published metabolic reconstruction, iCY1170_DHA, of the closely related *Schizochytrium limacinum* SR21 as a template [20].

Through iterative rounds of extensive manual curation and refinement, a high-quality GSM, termed iVS1191, was successfully reconstructed. iVS1191 showed significant improvements in both coverage and scope to that of iCY1170_DHA. The reconstructed model was phenotypically tested against experimental growth measurements in order to assess the validity of its predictions. These growth predictions were indicative of a questionable biomass composition, and highlighted the need for accurate and organism-specific biomass compositions to properly evaluate the predictive capabilities of the model.

The secondary aim of the project was to employ the reconstructed model to identify approaches for increased productivity of PUFA-containing lipids.

By applying computational methods for finding strain and genetic improvements, optimization strategies and potential bottlenecks for increased PUFA- and lipid production were identified. These strategies were in the form of double reaction mutants, whose target metabolite production rates were considerably greater than that of the wild type. Additionally, incorporation of transcriptomics data into the modeling framework was performed, elucidating the differential regulatory mechanisms that are associated with the metabolic shift from exponential growth to lipid accumulation.

Chapter 2

Background

Systems biology is a transdisciplinary field that aims at studying complex biological systems through the application of mathematical and computational techniques [21]. Whereas traditional biological research predominately has followed a reductionist approach, systems biology seek to understand the emergent properties of the biological system as a whole. This holistic approach has experienced a rapid and extensive development from its genesis almost two decades ago [21, 22]. The progress has predominately been driven by the increased availability of large-scale biomolecular data, which allows for a systems-level elucidation of biological processes [21].

Central to the field of systems biology is the use of mathematical models to represent and simulate biological systems. These partial representations of biological reality can offer new insight into the systems that they represent, and has generated considerable advancements in the understanding of biological phenomenon [23, 21].

The following section will therefore start by providing a thorough description of the underlying mathematical assumptions and principles of constraint-based metabolic modeling. To begin with, an introduction into linear programming is given, explaining its applicability in solving underdetermined systems of linear equations given an appropriate objective function. Subsequently, an overview of stoichiometric modeling and its application for modeling genome-wide metabolism using flux balance analysis (FBA) is presented. The following sections will then provide a comprehensive description of the process of GSM reconstruction and refinement, as well as the prediction of metabolic engineering strategies using this constraint-based modeling framework. Finally, the metabolism of lipid accumulation in oleaginous microorganisms is presented, providing an important reference point for the subsequent analyses performed on the reconstructed model.

2.1 Linear programming

Linear programming is a particular instance of mathematical optimization, and is widely used in a range of diverse areas such as economics, engineering, business and industry [24]. The wide applicability is a consequence of its mathematical formulation that can easily be used to model many real-life problems. In its most rudimentary form, the method aims at optimizing a linear objective function given a set of linear constraints [24, 25]. The general concepts of linear programming that follow are taken from the textbook "Optimization" by Lundgren et al. [25].

2.1.1 Linear program in standard form

A linear program can be broken down into three distinct components: an objective function consisting of a linear combination of decision variables, a set of constraints in the form of equalities or inequalities that determine the allowable solution space, and value restrictions on the decision variables that is to be determined [25]. The canonical form of a linear program is most often expressed as minimizing an objective function, while the constraints are in the form of equalities and the decision variables have to be non-negative. Constraints in the form of inequalities can easily be transformed to equalities, while negative or unrestricted variables may be converted to non-negative variables using appropriate variable substitutions. Additionally, maximization problems can be changed to minimization problems by simply minimizing the negative objective function of the original problem [25]. Consequently, non-standard formulations of linear programs can all be formulated using the following standard form

$$\min z = \sum_{j=1}^{n} c_j x_j \tag{2.1}$$

$$\text{subject to} \sum_{j=1}^{n} a_{ij} x_j = b_i, i = 1, ..., m \tag{2.2}$$

$$x_j \geq 0, j = 1, ..., n. \tag{2.3}$$

Here, z is the objective function that is to be minimized, x_j a decision variable, c_j the objective coefficient corresponding to the decision variable x_j, a_{ij} the constraint coefficient of x_j in the ith constraint, and b_i the right-hand side of the ith constraint. The goal of linear programming is to determine a set of values for the decision variables x_j such that z is minimized, while simultaneously not violating equations (2.2) and (2.3) [25].

The set of linear constraints in eq. (2.2) form a system of linear equations with n variables and m equations. Although the system may be solved algebraically whenever $n \geq m$, most problems appropriately modelled using linear programming are underdetermined (i.e. $n < m$). Here, the linear constraints form a solution space which subsist of all points that simultaneously satisfies all of these constraints (Figure 2.1). This region is known as a convex polyhedron; a region in n-dimensional space that is spanned by a convex set of points given by the intersecting linear constraints. The convexity of the solution space is a consequence of the linearity of the constraints, and causes any local optimum to

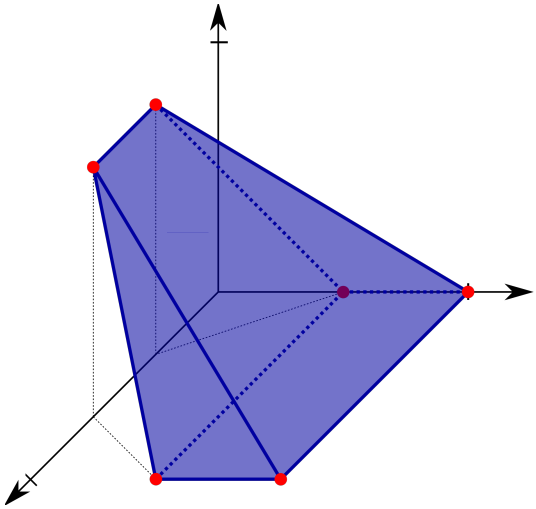


Figure 2.1: Feasible region in the form of a convex polyhedron in three-dimensional space. The solution space is determined by the intersecting constraints of the linear program. An optimal solution will reside in either of the indicated corner-points, or as a convex combination of a set of adjacent corner points (i.e. on a facet or edge of the convex polyhedron). Figure adapted from Ref. [26].

also constitute a global optimum. Given that the solution space is bounded and non-empty, along with the fundamental theorem of linear programming which states that an optimal solution will always reside on the boundary of this region, an optimal solution may be identified by searching for a local optimum along the edges of the polyhedron [25].

As the constraints span a solution space in $\mathbb{R}^n$, it is often convenient to algebraically express a linear program using the following matrix notation

$$\min z = \mathbf{c}^T \mathbf{x} \tag{2.4}$$

$$\text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.5}$$

$$\mathbf{x} \geq \mathbf{0}, \tag{2.6}$$

where $\mathbf{A}$ is an $m \times n$ coefficient matrix constituting all a_{ij} from eq. (2.2), $\mathbf{c}$ and $\mathbf{x}$ are $n \times 1$ vectors representing the objective coefficients and the corresponding decision variables, $\mathbf{b}$ is an $m \times 1$ vector representing the right-hand sides of the constraints, and $\mathbf{0}$ is the null vector. Consequently, $\mathbf{A}$ can therefore be regarded as a linear transformation, mapping elements of $\mathbb{R}^n$ onto elements of $\mathbb{R}^m$.

2.1.2 The simplex method

The most common approach for solving linear programs is the simplex method, an algorithm developed in 1947 by George Dantzig [27]. The simplex method relies on the already introduced notion that an optimal solution to a linear program will reside on the boundary of the feasible solution space in $\mathbb{R}^n$. More specifically, an optimal solution will

be located in a feasible corner-point solution given by the intersecting constraints. Although exhaustive enumeration of the objective value of all corner-points of the solution space would result in the identification of an optimal solution, this would be highly inefficient as the running time scales exponentially with the number of constraints. The simplex method exploits the fact that any local optimum will necessarily also be a global optimum due to the convexity of the solution space. This is employed in the criterion of optimality which identifies a local optimum by checking whether the objective value will improve in any of the adjacent corner points. If this is not the case, a local optimum has been reached, which consequently also must constitute a global optimum. The algorithm begins in a feasible corner-point solution and calculates whether the objective value in either of the adjacent corner-points is improving or not. It then iteratively moves between neighbouring corner-points in a direction of improving values of the objective function, such that the final optimal solution may efficiently be identified [25]. Although its worst case running time is exponential [28], its running time in practice turns out to be polynomial in the number of constraints [29]. In fact, for most conventional applications of linear programming, the simplex algorithm turns out to be as efficient as other solution algorithms with better worst-case time complexity, such as interior-point methods, as the simplex method rarely elicit its worst-case running time [30].

2.2 Metabolic modeling

The history of the mathematical representation of biochemical systems spans an entire century, and a multitude of diverse methods have been proposed in order to predict and explain the behavior of metabolic systems [31]. Even though modeling of *in vitro* enzyme kinetics using variants of Michaelis-Menten kinetics early on became an integral part of the field of biochemistry [32], the mathematical modeling of larger metabolic systems has not truly gathered traction until the past couple of decades [33]. A major driving force behind this shift has been an ever-increasing availability of large-scale biological omics data, whose integration with detailed biological knowledge from decades of research has laid the groundwork for high-quality models of metabolism [33]. The use of metabolic models also coincide aptly with the ethos of systems biology, where emergent properties of complex biological systems are elucidated through studying the systems as a whole [34]. It is therefore not a coincidence that systems biological research on biochemical systems entail a substantial emphasis on utilizing metabolic models to study the underlying biological mechanisms.

2.2.1 Stoichiometric modeling

The two major modeling frameworks of metabolism are dynamic and stoichiometric models, whose distinction has its roots in the model assumptions and mathematical formulations [35]. Dynamic models are used to simulate and predict the dynamic nature of biochemical systems using ordinary differential equations (ODEs). It usually begins with constructing ODEs which expresses the changes in concentration of the system components (i.e. metabolites, enzymes, regulators). The resulting set of equations are then solved, allowing insight into the changing behavior over time [36]. Central to dynamic

modeling is the concept of stability and robustness, which are investigated by examining the altered behavior of the system when parameter values or initial conditions are changed [37]. For this form of modeling, *a priori* knowledge of the biochemical topology and regulation is paramount. Additionally, extensive parametrization that either originates from experimental efforts or *in silico* parameter estimations is required [36]. This imposes substantial limitations on the applicability of this approach as the absence of high-quality kinetic parameters prevents the modeling of larger metabolic systems, or the metabolism of organisms of which little biochemical knowledge is established.

Stoichiometric models on the other hand, require minimal parametrization, and the only *a priori* knowledge called for is stoichiometric information on the metabolic transformations [38]. These models are based on a few key assumption which, in addition to negating the requirements for parametrization, has the enticing property that the resulting system of rate equations becomes linear [39].

The first assumption is that of mass balance; for every metabolite X_i in the metabolic system, a flux balance is given which mathematically is expressed as

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = \sum_{j=1}^{n} s_{ij} v_j \tag{2.7}$$

Here, s_{ij} is the stoichiometric coefficient of metabolite i in reaction j, and v_j the corresponding reaction flux. Eq. (2.7) states that the change in concentration of any metabolite X_i is given as the sum of the stoichiometrically weighted fluxes of the reactions where X_i is a constituent.

Secondly, as the time constants which describe transient metabolic variations are much smaller than the time constants of cell growth, the time derivatives may be ignored by investigating the steady-state behavior of the system [40]. Consequently, eq. (2.7) is transformed to the following homogeneous system

$$\sum_{j=1}^{n} s_{ij} v_j = 0, \tag{2.8}$$

or similarly, in matrix notation

$$\mathbf{S}\mathbf{v} = \mathbf{0}, \tag{2.9}$$

where $\mathbf{S}$ has the generic form

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \dots & s_{mn} \end{bmatrix}. \tag{2.10}$$

For m metabolites and n reactions, $\mathbf{S}$ is an $m \times n$ stoichiometric matrix whose entries s_{ij} are the stoichiometric coefficients of metabolite i in reaction j, while $\mathbf{v}$ is an $n \times 1$ flux vector. By solving this homogeneous system of linear equations, the flux distribution of the entire metabolic network of interest may be predicted [39]. As aforementioned earlier and aptly stated in eq. (2.9), the only necessary knowledge required is the stoichiometric

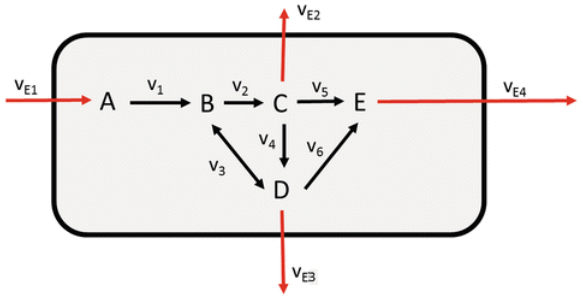


Figure 2.2: Toy example of a metabolic network containing 5 metabolites (A - E), and 10 metabolic reactions. The latter of which consist of 4 exchange reactions (v_{E1} - v_{E4}) importing or exporting metabolites to/from the system, while 6 are intracellular metabolic transformations (v_1 - v_6). Figure adapted from Ref. [39].

relationships between all the constituent metabolites in the metabolic system of interest. The enticing idea of modeling the genome-scale metabolism of an organism is therefore only restricted by the amount of available biochemical information [18].

To illustrate the essence of stoichiometric modeling, consider the following example of a minor metabolic network containing 5 metabolites and 10 reactions, as depicted in Figure 2.2. The exchange reactions v_{E1} - v_{E4} represent the input and outputs of the metabolic system. The imported metabolite A is converted through a series of interconnected reactions into either of the 4 other internal metabolites B - E. These interconversions are given by the 6 internal reactions v_1 - v_6. Assuming that the fluxes (i.e. reaction rates) are given by the reaction names, flux balances may be expressed for every metabolite in which producing and consuming reactions are given positive (+) and negative (-) signs, respectively. The mathematical representation is given as

$$\frac{dA}{dt} = v_{E1} - v_1 \tag{2.11}$$

$$\frac{dB}{dt} = v_1 - v_2 - v_3 \tag{2.12}$$

$$\frac{dC}{dt} = v_2 - v_4 - v_5 - v_{E2} \tag{2.13}$$

$$\frac{dD}{dt} = v_3 + v_4 - v_6 - v_{E3} \tag{2.14}$$

$$\frac{dE}{dt} = v_5 + v_6 - v_{E4}. \tag{2.15}$$

These balance equations can concisely be formulated in the form of a stoichiometric matrix $\mathbf{S}$ given by

$$\mathbf{S} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & -1 \end{bmatrix},$$

where the order of the rows are according to the metabolites A - E, and the columns according to the reactions v_1 - v_6, v_{E1} - v_{E4}.

Finally, due to the assumption of a metabolic steady-state, the unknown fluxes can be calculated by solving the resulting system of linear equations in eq. (2.9). Through the underlying assumptions of stoichiometric modeling, the unknown rate equations have now been replaced by the placeholder values v_j, whose value directly denotes the metabolic flux through reaction j.

2.2.2 Flux balance analysis - FBA

For most real-life biochemical systems, the number of reactions exceeds the number of metabolites [19]. As a consequence, the $m \times n$ stoichiometric matrix $\mathbf{S}$ contains more columns than rows ($n > m$). The associated system of homogeneous linear equations of the form seen in eq. (2.9) thus denotes an underdetermined system of equations, which cannot be solved using Gaussian elimination alone. In fact, any solution to the system of linear equations reside within the null space of $\mathbf{S}$ [19]. Several methods have been developed to solve these forms of problems [41]. However, owing to its simple formulation and efficient solvability, FBA has become the principal methodology to predict the flux distributions of metabolic networks modeled using a stoichiometric framework [19].

FBA applies the methodology of linear programming to identify a solution within the aforementioned null space. The null space can be thought of as the feasible solution space to the system of linear equations in which an optimal solution resides. By optimizing an evolutionary motivated objective function, this optimal solution will encompass a more biologically realistic flux distribution [19]. In addition to the linear constraints given by the stoichiometric matrix, additional constraints in the form of upper or lower bounds are usually added on particular fluxes, distinguishing the sets of reversible and irreversible reactions [42]. By doing this, the unbounded null space is transformed to a bounded hyperspace which, because of the linearity of the constraints, forms a convex polyhedron (Figure 2.3). An optimal solution will consequently reside in an intersection of the given constraints, and may therefore be determined using the simplex method or other solution methods of linear programs [19].

The general FBA problem can therefore be stated as the following linear program

$$\max z = \mathbf{c}^T \mathbf{v} \tag{2.16}$$

$$\text{s.t.} \quad \mathbf{S}\mathbf{v} = \mathbf{0} \tag{2.9}$$

$$\mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub}, \tag{2.17}$$

where $\mathbf{c}$ is the $n \times 1$ objective coefficient vector, which determines the objective function as a linear combination of the reaction fluxes given by the $n \times 1$ flux vector $\mathbf{v}$. $\mathbf{S}$ is the $m \times n$ stoichiometric matrix denoting all biochemical transformations of the metabolic system. Eq. (2.9) represents the homogeneous system of linear equations, while eq. (2.17) specifies the upper ($\mathbf{lb}$) and lower ($\mathbf{lb}$) bounds on the flux variables. Other forms of restrictions of physiochemical origin may also be added to further restrain the allowable solution space [19]. The formulation of these constraints reduces the size of the feasible solution space and allows for more accurate predictions of metabolic fluxes [42].

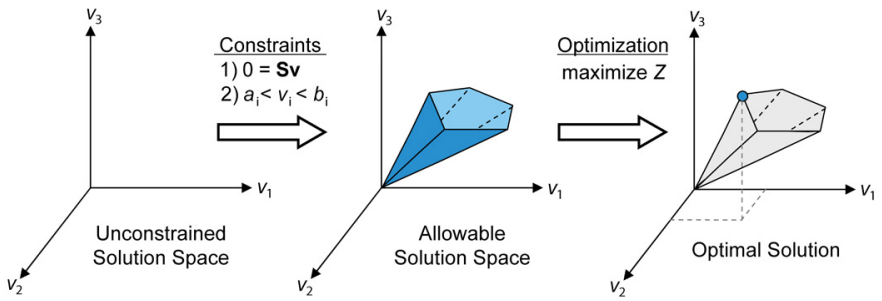


Figure 2.3: Illustrative example of the FBA problem. 1) The stoichiometric constraints, and 2) flux variable bounds impose restrictions on the unconstrained solution space, reducing the range of feasible flux distributions. The resulting convex polyhedron is treated as the feasible solution space in a linear program in which an objective function composed of a linear combination of reaction fluxes is optimized. Figure taken from Ref. [18].

The choice of objective function poses a difficult challenge in the FBA formulation. For microorganisms, the assumed objective function is usually cellular growth, which rest on the assumption that cells that maximizes growth tend to outcompete the other cells in the population [43]. Although several other objective have been proposed, the optimization of biomass production seems to fit well with experimental data, especially when predicting the cellular growth of exponentially growing microorganisms [43, 44]. The real cellular objective is most likely a combination of several interdependent goals, and will presumably change during various growth phases, metabolic demands, and environmental conditions [43].

2.2.3 Flux variability analysis - FVA

Although the optimization of an objective function allows for the identification of a solution from a larger solution space, this solution is often non-unique. In many cases, several alternative flux distributions give rise to the same optimal objective value, indicative of a highly flexible metabolic system that is able to obtain the same optimal phenotype using a range of different flux distributions [45]. This variability of metabolic fluxes can be calculated with an approach called flux variability analysis (FVA), which calculates the lower and upper bounds of every reaction flux in the model [45]. This is expressed by the two linear programs

$$\max / \min v_j \tag{2.18}$$

$$\text{s.t.} \quad \mathbf{Sv} = \mathbf{0} \tag{2.9}$$

$$\mathbf{c}^T\mathbf{v} \geq \gamma z \tag{2.19}$$

$$\mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub} \tag{2.17}$$

Here, the reaction flux v_j is maximized and minimized to obtain its upper and lower bounds given the stoichiometric constraint of eq. (2.9) and the flux boundary constraints of eq. (2.17). The last constraint, eq. (2.19), expresses the minimal objective value that

has to be obtained, which for optimal phenotypes is given when $\gamma = 1$, while sub-optimal objective states can be explored when $0 \leq \gamma < 1$.

2.2.4 Minimization of metabolic adjustment - MOMA

The calculations of metabolic flux distributions by optimizing biomass production using FBA is based on the evolutionary motivated assumption that the organism has evolved to reach an optimal growth performance [19, 46]. However, following a genetic perturbation such as a gene knockout, this assumption falls short as the perturbed metabolic network of the organism has not been subject to the same evolutionary pressure to reach optimality. Minimization of metabolic adjustment (MOMA) is a method which tries to reconcile this problem when calculating the flux distribution of a perturbed metabolic network (e.g. knockout mutant) [46]. It relies on a separate hypothesis where it is assumed that the metabolic fluxes of the mutant network is minimally different from that of the unperturbed network, thus minimizing the metabolic adjustments from the wild type metabolic phenotype. This can mathematically be explained as finding a feasible solution point in the *n* dimensional flux space for the mutant network that is closest in Euclidean distance to the optimal solution of the wild type network (Figure 2.4).

The solution to the MOMA problem of a reaction knockout of reaction j, *j'*, is a flux vector $\mathbf{x}$ within the knockout solution space $\Phi^{j'}$ that has a minimal Euclidean distance to the wild type optimal solution $\mathbf{v}$ within the wild type feasible solution space Φ. This Euclidean distance can mathematically be expressed as

$$D(\mathbf{v}, \mathbf{x}) = \sqrt{\sum_{j=1}^{n} (v_j - x_j)^2}, \tag{2.20}$$

where v_j and x_j are the fluxes of reaction j in the wild type and mutant network, respectively. Finding a flux vector $\mathbf{x} \in \Phi^{j'}$ thus correspond to a minimization of this distance D. Due to the non-linearity of the objective function D, quadratic programming has to be applied to be able to solve the MOMA problem. Quadratic programming is similar to linear programming in that the constraints are linear, but dissimilar in that the objective function is quadratic [25]. The general objective function $f(x)$ for a quadratic program can be expressed as

$$f(x) = \mathbf{L}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}. \tag{2.21}$$

Here, $\mathbf{x}^T$ denotes the transpose of $\mathbf{x}$, the $n \times 1$ vector $\mathbf{L}$ and $n \times n$ matrix $\mathbf{Q}$ denotes the linear and quadratic part of the objective function, respectively. The distance D can be transformed to this form by noting that minimizing D is equivalent to minimizing its square, and further setting $\mathbf{Q}$ to be the identity matrix $\mathbf{I}$, and letting $\mathbf{L} = -\mathbf{v}$. The MOMA problem for a knockout of reaction *j'* can consequently be expressed as

$$\min f(x) = \frac{1}{2}\mathbf{x}\mathbf{I}\mathbf{x}^T + (-\mathbf{v})\mathbf{x} \tag{2.22}$$

$$\text{s.t.} \quad \mathbf{S}\mathbf{x} = \mathbf{0} \tag{2.23}$$

$$v'_j = 0 \tag{2.24}$$

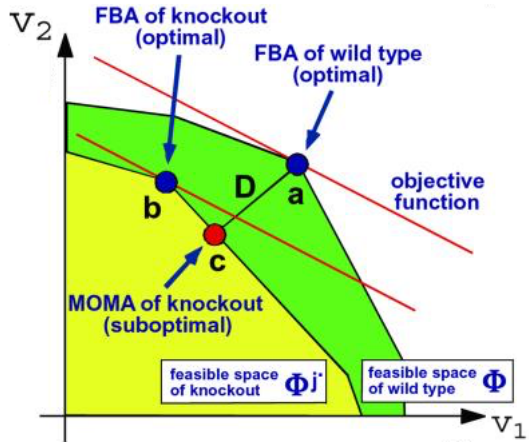


Figure 2.4: The criteria of optimality for the FBA and MOMA problem mapped onto a two-dimensional flux space for clarity. The solution to the MOMA problem of the perturbed network is a feasible solution $\mathbf{x} \in \Phi^{j'}$ (denoted as the point **c**) that resides closest to the optimal solution of the wild type network, $\mathbf{v} \in \Phi$ (given as **a**). MOMA finds this solution by minimizing the Euclidean distance **D** using quadratic programming. Also indicated is the optimal solution of the FBA problem for the knockout mutant which assumes evolutionary optimized growth performance (point **b**). Figure adapted from Ref. [46].

$$\mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub}, \tag{2.17}$$

where the flux through the knocked out reaction v'_j is constrained to zero in eq. (2.24) [46].

2.3 Genome-scale metabolic models - GSMs

To properly understand any biochemical system, one needs to discern its role and functionality within a larger framework of genome-scale metabolism [47]. Although biochemical literature historically have identified biochemical modules which serve some particular function as unique entities, they are in fact components of a highly interconnected network of biochemical conversions [48]. Their behavior are constrained and regulated by a complex web of interacting metabolites and enzymes. As such, their function might not be easily defined, and may change as a consequence of varying metabolic demands and environmental conditions [49]. GSMs acknowledges this fact by taking into account the entire metabolism of an organism to enable flux predictions of all reaction rates throughout the metabolic network [50]. This is accomplished by concisely representing the totality of metabolic activities within a stoichiometric framework. These activities are deduced using the available genomic, physiological and bibliomic data on the organism of interest to infer all metabolic capabilities.

In addition to flux simulations, these models provide direct correlations between the genetic information and the biomolecular phenotype through gene-reaction relationships [18]. The effects of genetic interventions can therefore be predicted, thus enabling the

guiding of metabolic engineering strategies [51]. Additionally, as these models encompass every enzymatic capability of an organism, they are also regarded as excellent repositories of species-specific biochemical information [42].

2.3.1 Reconstruction of GSMs

The reconstruction of GSMs is a highly labour-intensive procedure which can be partitioned into four major stages, with a subsequent fifth and final stage; employing the model for phenotypic predictions. These stages are further subdivided into 96 steps, as described in Ref. [18] (Figure 2.5). Although consecutive in order, these stages are organized in an iterative manner in which additional model curation is performed such that the predictions progressively match the expected biochemical phenotype. While many of the steps require, or at least benefit from manual curation, several steps have become partly, if not fully automated. These steps are performed using computational software and algorithms which drastically reduces the time spent on model reconstruction [52]. The following section gives an overview of these five stages, and unless stated otherwise, is taken from [18].

Stage 1 - Draft reconstruction

The first stage involves the initial generation of a draft reconstruction. Here, appropriate metabolic functions are added to the model in the form of a stoichiometric matrix based on the enzymatic capabilities of the cell. The primary source of this information resides in the genome annotation of the organism of interest. The functionality of the gene product found in the annotation are either predicted using homology-based methodologies or through evidence from experimental efforts. The homology-based methods involve searching a range of reference databases for significantly similar genes or proteins, and by doing so, deducing their putative functionalities. This approach is the norm for organisms of which little biological insight is available, while the deduction of functionality from experimental efforts is mainly the case for extensively studied model organisms. Irrespective of the methodology, the resulting metabolic functionalities are gathered as a collection of biochemical reactions along with the associated genes. These form the initial draft model and provide a primary scaffold where additional reactions, metabolites, and genes can be added later on in the reconstruction process.

Through the recent years, several computational procedures have been created in order to facilitate an automatic generation of these draft reconstructions [52]. In general, these methods can be subdivided into two distinct groups. The first group of methods utilizes already constructed GSMs in combination with putative genetic homology to infer which reactions a particular gene should be associated with [53, 54], while the second group employs large biochemical databases, where metabolic reactions are annotated to genes from a taxonomically diverse range of organisms [53, 55]. In both methodologies, significantly similar genes are assumed to be homologous, and the reactions associated with the genes in either the template model or biochemical database are subsequently incorporated into the new draft reconstruction, greatly assisting in the construction of the initial metabolic model [53]. A prerequisite of these methodologies is to select appropriate parameter values for the sequence comparisons. Closely related organisms require higher significance

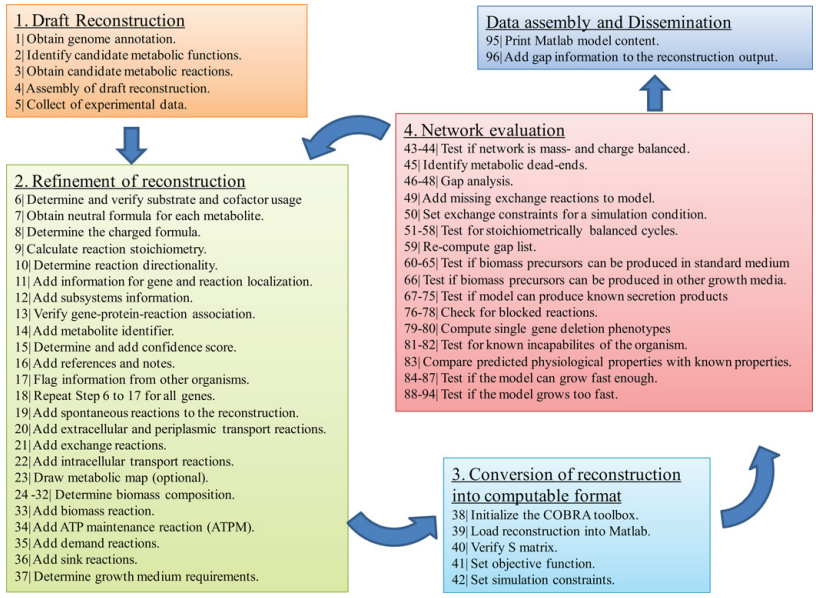


Figure 2.5: The five stages of GSM reconstruction, where the stages are subdivided into a total of 96 unique steps. The iterative nature of the model reconstruction procedure in steps 2-4 is indicated by the blue arrows. Figure taken from Ref. [18].

levels to ascertain genetic homology, while more distantly related organisms call for less stringent similarities. At the same time, the evolutionary relatedness of the organisms are also of importance, as the probability that the inferred homologous genes actually carry out the same metabolic function are inversely proportional to the taxonomic distance [56]. Consequently, as the parameters and taxonomic relation between the organisms are interdependent of one another, a lot of care has to be taken in order to generate a draft model of satisfactory quality.

Stage 2 - Refinement of the reconstruction

The second stage encompasses the refinement and curation of the draft reconstruction, which for the most part require considerable manual inspection and verification of the model components. This is especially the case for the automatically generated draft reconstructions where the resulting gene functionalities might be erroneous. Whereas both the sequences and activity of enzymes associated with the central carbon metabolism are highly conserved, the functionality of enzymes in more peripheral or organism-specific metabolic pathways are often harder to pinpoint. Consequently, particular care has to be made in the inference of the metabolic capabilities of these proteins, either to prevent the inclusion of incorrect metabolic functions, or on the other hand, leave out essential enzymatic activity.

Even though the majority of biochemical transformations require enzymes in order to occur at a satisfactory rate, certain reactions can take place without the assistance of en-

zyme catalysis [57]. The exclusion of these spontaneous reactions may subsequently lead to gaps in the metabolic network. The elucidation of these reactions is therefore a highly important step in the process of model refinement. Special emphasize should also be given on correctly representing the gene-reaction associations so as to accurately describe the relation between the encoded protein and the reaction. These relationships are expressed as boolean rules in which constituents of enzymatic complexes and isozymes are associated via 'and'- and 'or'-logic, respectively. These relations links a gene with the appropriate chemical reactions of the encoded enzyme which, because of the boolean formulation, allows for *in silico* gene essentiality predictions [58].

As mass balance is a central concept of the stoichiometric modeling framework, it is necessary to correctly define appropriate metabolite formulas so as the resulting biochemical reactions are mass balanced. Although the neutral metabolite formulas are easily obtainable, the charged formulas are dependent on the pH of the associated cellular compartment and may be more difficult to ascertain. Certain biochemical databases such as MetaCyc do contain charged formulas [59], but in some cases it might be more appropriate to estimate the protonation state when the intracellular pH levels are available. When the mass and charge of every metabolite has been established, the proper metabolite stoichiometry for every reaction may be surmised.

The biochemistry of eukaryotic organisms is characterized by the spanning of several subcellular compartments [60]. This compartmentalization causes the subcellular sections to fulfill particular metabolic roles by housing entire or subsets of metabolic pathways, providing a major driving-force behind the characteristic behavior of eukaryotic metabolism [61]. The subcellular targeting of cellular proteins to these compartments is an intricate and tightly regulated process. Central to this mechanism are signal peptides in the synthesized protein that carry information of where the protein should be localized [62]. Several computational algorithms have been developed in order to identify these signal sequences, and by that, predict the most likely subcellular localization that a given protein will reside in [63]. Even though certain proteins are localized to multiple compartments [64], most of these subcellular predictors only return a unique subcellular prediction [63]. Consequently, proteins targeted to multiple compartments are quite a challenge to identify and may constitute a major source of gaps in the metabolic networks of eukaryotic GSMs.

The transport of metabolites across these subcellular compartments are generally carried out by specific transporter proteins [65]. These proteins possess highly specific affinities for sets of metabolites with similar molecular structures. The identification of transporter proteins from predicted protein sequences are quite straightforward as these proteins are evolutionary highly conserved, owing to the existence of universal membrane-spanning motifs [65]. However, the inference of which subset of metabolites a particular transporter actually translocates across the membrane, and by what molecular mechanism, is more of a challenge to ascertain [66]. Consequently, considerable care is paramount when adding these reactions to the model, both to prevent the inclusion of invalid reactions, as well as the exclusion of necessary metabolite transport.

Associated exchange reactions also have to be added to the model. These pseudo-reactions define the boundaries of the metabolic system and determine which metabolites are able to enter and leave the system. The exchange reactions are added as extracellular reactions, which subsequently also determine the growth medium of the model. By

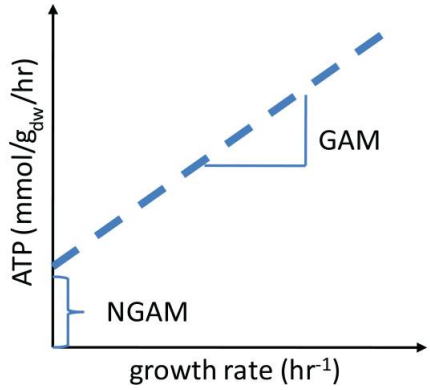


Figure 2.6: Derivation of GAM and NGAM using growth data from chemostat experiments in which the specific growth rates are plotted against the ATP consumption rates. Figure taken from Ref. [18].

altering the associated exchange rates, growth on particular nutrient compositions can be simulated, and the model predictions can be correlated with the corresponding experimental phenotypes.

To enable the prediction of growth rates, an abstractive biomass reaction is added to the model. Here, the substrates consist of all metabolic precursors required for the species-specific biomass production [67]. These usually include amino acids, nucleotides, lipids, carbohydrates, vitamins and cofactors. Each of these precursors are stoichiometrically weighted such that the generated biomass has a molecular weight of 1 g mmol^{-1}, thus enabling quantitative predictions of specific growth rates (h^{-1}) [68]. In order to accurately simulate growth rates in accordance with experimentally determined growth rates, the quantitative biomass composition of the organism has to be determined experimentally [69]. Additionally, the energetic demand related to growth-associated processes such as the polymerization of the cellular macromolecules (i.e. DNA, RNA and proteins) is included in the biomass reaction in the form of ATP hydrolysis (termed growth-associated maintenance (GAM)). Non-growth associated maintenance (NGAM) is also represented as an independent ATP consumption reaction. The derivation of the associated stoichiometric coefficients are usually performed by linear regression of growth data from chemostat experiments (Figure 2.6) [67].

Stage 3 - Conversion of reconstruction into computable format

The third, and shortest stage involves the conversion of the draft reconstruction into a computational format. Several computational environments exist for this purpose, such as the COBRA toolbox [70], CellNetAnalyzer [71], and the RAVEN Toolbox [54]. These provide functionalities for analyzing the predictive properties of the model, and can be used to facilitate the labour-intensive process of model refinement. Additional functionalities are commonly incorporated within these frameworks, allowing for the employment of a wide range of computational techniques on the metabolic models, as well as greatly assisting in their generation [70].

Stage 4 - Network evaluation

The fourth stage can be regarded as the debugging phase of the model reconstruction process. Here, extensive evaluation of the model properties are performed so as to reduce the likelihood of incorrect model predictions. One of the most important tests is that of mass and charge balance. To prevent the loss or creation of mass, which can cause erroneous flux predictions, all the reactions have to be stoichiometrically balanced based on the elemental composition of the metabolic constituents. The most common issue is the inclusion or omission of protons which, based on the chosen protonation state of the metabolites, either has to be removed or added to the associated reaction.

An inevitable property of draft reconstructions is the existence of gaps in the metabolic network. These gaps result in blocked reactions and entire pathways which cannot carry any flux due to the model assumption of mass balance and a metabolic steady state [72, 73]. Consequently, it is often the case that the initial model reconstruction cannot grow as one or multiple biomass components are unable to be produced. These missing metabolic functions can be a result of unknown enzymatic reactions, the existence of promiscuous enzymes, or insufficient coverage of the genome annotation [72]. Irrespective of why a gap might be present, considerable efforts are necessary to fill these, allowing for more realistic flux predictions.

As with the initial draft reconstruction stage, a broad selection of automatic procedures exist which can be used to generate gap filling candidates for a draft reconstruction [74, 75, 76]. These operate by altering the network connectivity through providing a list of candidate reactions, whose addition to the model will fill metabolic gaps. Some algorithms also generate putative gene-assigments, which may assist in the identification of the associated genes [73]. The quality of these gap-filling reactions is however a matter of debate [77] and substantial manual verification of these suggestions has to be performed in order to prevent issues like over-fitting by adding reactions that do not occur *in vivo* [72]. Manual inspection of the gaps is also a much used strategy in order to fill the holes in the metabolic network. Although time consuming, it generally leads to a more realistic set of candidate reactions by preventing the inclusion of erroneous reactions that may be added by the algorithmic strategies [77]. These candidate reactions can be identified by investigating the metabolic surrounding of the gaps, which can be found in biochemical pathway databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) and MetaCyc [78, 59]. Subsequently, the proposed reactions can then be evaluated by inspecting whether there exists genomic evidence for the existence of the associated enzymes, or whether other physiological evidence might indicate their presence [77].

Another central task of the network evaluation stage is to compare the model predictions with experimental data on the biomolecular physiology. Validating growth/non-growth on a wide range of nutrient compositions is an efficient strategy in order to evaluate the quality of the reconstructed model. Growth rate comparisons is also a commonly used procedure in which the simulated growth rate of the model is compared with the actual growth rate of the organism. A prerequisite for this is measured uptake rates of the limiting nutrient, such that the availability of the growth-limiting metabolic precursor is the same for both the organism *in vivo* and the model *in silico*. Too low growth rates indicate that one or several biomass precursors cannot be synthesized at a sufficient rate, which suggests that some particular metabolite is growth-limiting. On the other hand, too high

growth rates could suggest that the estimated GAM or NGAM might be insufficient, or that incorrect reactions have been added to the metabolic network. It could also indicate that the optimization of biomass production might be an unsuitable objective function.

A common issue with compartmentalized GSMs is the presence of thermodynamiccaly infeasible energy-generating cycles (EGCs). These subgroups of type II extreme pathways can give rise to spontaneous generation of reducing power which could be used to drive ATP synthesis without any input of matter [79]. These pathways are biochemical loops in which matter is transformed from an initial state through a series of interconnected biochemical reactions back into the same state in which it started [80]. Identifying and removing these cycles are essential in order to accurately predict growth rates and associated flux distributions.

Stage 5 - Draft assembly and dissemination

The fifth and final stage consists of the model employment to achieve novel insight into the genome-scale metabolism of the organism. An extensive amount of functionalities within the various constraint-based computational frameworks can be employed to utilize the predictive power of GSMs [70]. The metabolic fluxes can be estimated using FBA, where an appropriate objective function is chosen to be optimized (e.g. biomass production) [19]. The calculated fluxes can subsequently be used to obtain new understanding into how the flow of matter are partitioned onto the biochemical network. Effects of genomic interventions such as gene deletion, over-expression and dampening strategies, and additions of novel biochemical reactions can efficiently be simulated before any experimental efforts are initiated. Thus, these GSMs can be regarded as particularly helpful tools in the generation of biologically realistic hypotheses on the metabolic phenotype [33].

2.3.2 Employment of GSMs for the prediction of metabolic engineering strategies

The field of metabolic engineering strives to exploit the large biochemical potential of metabolic networks in order to improve the biomolecular phenotype and productivity of microorganisms [40]. Most often, the goal is to identify strategies to reroute the metabolic fluxes towards particular pathways in order to increase the yield and production rate of a metabolite of interest [51]. Through genetic interventions, the flux partitioning of the metabolic network is altered so as to increase the productivity. These interventions can either adjust or remove the activity of existing enzymes and/or regulators, or introduce novel metabolic capabilities through the expression of heterologous genes [81]. Historically, these modifications were informed by extensive knowledge of the underlying metabolic organization and regulation [82]. However, the recent employment of metabolic models has turned out to be a valuable mode of research, both with regards to the simulation of these genetic interventions, as well as the generation of novel overproduction strategies using the modeling frameworks [51].

As the genome-wide metabolism consists of a highly interconnected network of biochemical reactions, genetic interventions at seemingly distant parts of the metabolic network may greatly affect the production rate of a particular compound. The identification and effects of these apparently ”counter-intuitive” strategies are therefore often very hard

to deduce without the use of *in silico* flux simulations [83]. Consequently, constraint-based metabolic models such as GSMs have been widely used in combination with FBA-related methods to generate strategies for increased productions of metabolites of interest. These approaches propose sets of genetic interventions that enables the metabolite of interest to be produced at higher levels compared to a predefined wild type state [84]. Several of these methods operate by altering the connectivity of the metabolic network using gene deletions such that the flux of the metabolite-producing reaction and the cellular objective (e.g. growth), become coupled. By optimizing for the same cellular objective, this coupling forces flux through the target reactions, thus leading to a higher production rate of the metabolite of interest [51]. These approaches have led to the elucidation of several successful genetic interventions, such as increased lycopene and succinate production in *Escherichia coli* [85, 86], and reduced glycerol production with a concurrent increased yield of ethanol in *Saccharomyces cerevisiae* [87]

2.4 Lipid accumulation in oleaginous microorganisms

The metabolic organization and behavior of lipid biosynthesis and accumulation is highly conserved among oleaginous microorganisms [16]. Central to the initiation of lipid accumulation is the depletion of an essential nutrient (e.g. nitrogen), while a steady supply of carbon is maintained. The cell halts its growth and redistributes its metabolic fluxes from biomass formation into the biosynthesis of lipid precursors (Figure 2.7). The synthesized lipids are stored as intracellular lipid droplets, which function as energy reserves during the ensuing period of nitrogen starvation [16]. One of the earliest responses to nitrogen deprivation is the enhanced activity of AMP (adenosine-5'-monophosphate) deaminase (AMPD, EC:3.5.4.6), which catalyzes the irreversible deamination of AMP, generating the much needed ammonium ($NH_4{}^+$)

$$\mathrm{AMP} + \mathrm{H_2O} \longrightarrow \mathrm{NH_4}^+ + \mathrm{IMP}. \tag{2.25}$$

This increased activity causes an intracellular depletion of AMP, which, due to its role as an obligatory allosteric activator of the mitochondrial isocitrate dehydrogenase, reduces the extent of isocitrate oxidation in the tricarboxylic acid (TCA) cycle [16]. Through rapid equilibration between isocitrate and citrate, accumulation of citrate ensues, which subsequently gets exported out of the mitochondria, primarily via an antiport-mechanism along with malate. The cytosolic citrate is then cleaved by ATP citrate lyase (ACL, EC:2.3.3.8), forming oxaloacetate and the central metabolite acetyl coenzyme A (CoA). The latter is subsequently carboxylated by acetyl-CoA carboxylase (ACC1, EC:6.4.1.2), forming the immediate precursor of fatty acids; malonyl-CoA. Malonyl-CoA is then used in the successive elongation of a wide range of fatty acid moieties, depending on the species [16].

In thraustochytrids, two distinct pathways are utilized: the classical fatty acid synthase (FAS) pathway, and a polyketide synthase (PKS) pathway. The FAS pathway is utilized in the biosynthesis of saturated fatty acid of chain lengths C16 - C18, while longer PUFAs such as DHA and eicosapentaenoic acid (EPA), are synthesized via the PKS pathway [88]. The latter pathway operates by retaining the unsaturation of certain acyl-intermediates

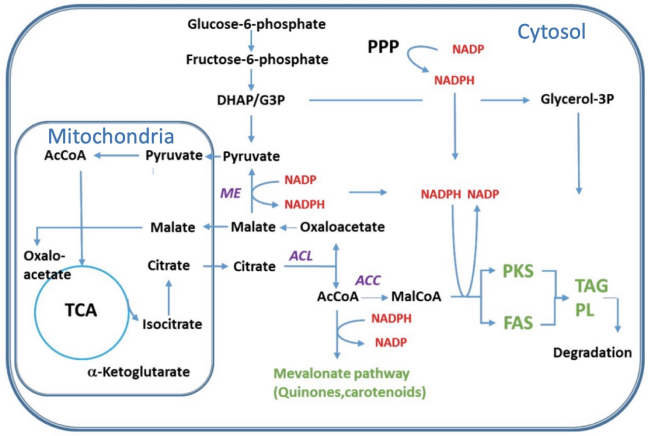


Figure 2.7: Overview of the metabolic pathways involved in the production of the fatty acid precursor malonyl-CoA, as well as reducing power in the form of NADPH. Depletion of nitrogen initiates a metabolic cascade in which citrate accumulates in the mitochondria due to the reduced activity of isocitrate dehydrogenase. Citrate is then exported to the cytosol, where it generates acetyl-CoA, which subsequently gets carboxylated, forming malonyl-CoA. Malonyl-CoA is then shuttled into either of two pathways of fatty acid biosynthesis: the traditional fatty acid synthase (*FAS*) pathway or a polyketide synthase (*PKS*) system. The reducing power are thought to be generated from the pentose phosphate pathway (PPP) or malic enzyme (*ME*). The generated fatty acids are further used, along with glycerol-3-phosphate, to create TAGs and phospholipids (PLs). The former are stored as intracellular lipid droplets and constitute the majority of the generated lipids. Abbreviations: AcCoA: acetyl coenzyme A, TCA: tricarboxylic acid cycle, *ACL*: ATP-citrate lyase, *ACC*: acetyl-CoA carboxylase.

through the iterative elongation, thus lowering the molar demand for reducing power in comparison with the classical elongation/desaturation scheme (Figure 2.8) [16]. The reducing power required for the biosynthesis of these fatty acids is hypothesized to be generated via two major pathways: the NADPH-generating reactions of the pentose phosphate pathway (PPP), and the activity of cytosolic malic enzyme (ME, EC:1.1.1.40) [16]. The latter catalyzes the oxidative decarboxylation of malate to pyruvate, forming NADPH in the process. This malate is thought to be created from the oxidation of oxaloacetate, the secondary product of ACL [89]. Although the extent to which these two pathways contributes to the NADPH-generation in thraustochytrid is unclear, it is regarded as one of the primary bottlenecks of fatty acid biosynthesis, along with the generation of malonyl-CoA [16].

In *A.* sp. T66, the majority of biosynthesized fatty acids are esterified along with glycerol-3-phosphate to form TAGs during the lipid accumulation phase [90]. The regulation of the the fatty acid composition of these lipids are poorly understood, and whether the biosynthetic enzymes have varying substrate preferences or whether degradation of certain classes of fatty acids are more predominant than others is unclear. Similarly, the regulatory mechanisms behind the metabolic shift from biomass production during exponential growth to lipid accumulation following nutrient depletion is also fairly unsettled

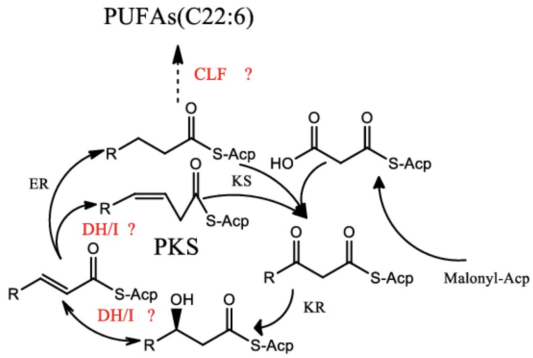


Figure 2.8: Putative PKS pathway of PUFA biosynthesis in thraustochytrids. The acyl-chain is successively elongated through the condensation of 3-ketoacyl-ACP and malonyl-ACP by ketoacyl synthase (KS). The subsequent reduction by ketoacyl reductase (KR) and dehydration by dehydratase (DH) generates a *trans*-enoyl-ACP intermediate, which either may be isomerized to the *cis*-isomer by a proposed isomerase domain (I), or reduced to form a saturated acyl-ACP by enoyl reductase (ER). By retaining the unsaturated bonds during the biosynthetic process, less reducing power is needed by the cell to synthesize these PUFAs. Taken from Ref. [91].

[90].

Chapter 3

Software and methods

The following chapter will give a detailed description of the software and various methods employed throughout the project. To begin with, a succinct introduction to the software that was used during the model reconstruction is given. The subsequent section will then provide a comprehensive review of the model reconstruction, refinement and validation. Finally, the ways in which the final reconstruction was employed for novel phenotypic predictions is described.

3.1 Software

3.1.1 MATLAB

MATLAB is a computational environment and scripting language developed by MathWorks specifically tailored for numerical analysis [92]. With a particular emphasis on matrix manipulations, the programming language is widely used within the engineering disciplines, as well as the field of applied mathematics. Extensions to MATLAB exist as separate toolboxes with specialized functionalities, either developed by MathWorks or as independent, community-generated software suites. During this project MATLAB 2017b was used for the model reconstruction and refinement, as well as the subsequent model employment [92].

3.1.2 Python

The pre-processing of the JSON format files for constructing metabolic maps in Escher were performed using Python version 3.6.4 [93].

3.1.3 COBRA toolbox

The COBRA (Constraint-based reconstruction and analysis) toolbox is a comprehensive repository of software for the refinement and analysis of constraint-based metabolic mod-

els [70]. The toolbox is community-generated and interoperable, allowing for the efficient utilization of a wide range of computational approaches on these models. For the main refinement and analysis of the reconstructed GSM, the COBRA toolbox v3.0 was used within the MATLAB computational environment [70]. COBRApy v0.13.4, an object-oriented COBRA package for Python, was also utilized for the aforementioned preprocessing of the JSON format files [94].

3.1.4 RAVEN Toolbox

The RAVEN (Reconstruction, analysis, and visualization of metabolic networks) Toolbox is a standalone MATLAB toolbox that enables the semi-automatic reconstruction and analysis of GSMs [53]. It proposes two modes of action in order to generate a draft reconstruction: (1) using one or multiple template models and protein homology to ascertain which reactions are to be selected and added to the draft reconstruction, or (2) identifying homologous proteins in the biochemical database KEGG, and subsequently creating a model using the associated reactions. RAVEN Toolbox v1.0 was employed to generate two separate draft models using both of these approaches [53].

3.1.5 ModelExplorer

ModelExplorer is a software program used to identify inconsistencies within the metabolic network of GSMs. The software visualizes metabolic networks as bipartite graphs of interconnected nodes of metabolites and associated reactions. The nodes are highlighted so as to indicate whether the associated reactions are able to carry any flux or not. Using various inconsistency checks, the origins of these blocked reactions can be identified, markedly assisting in the labour-intensive manual curation of the metabolic model [95]. The ModelExplorer software was used during the various gap-filling stages of the model reconstruction, as well as in identifying the sources of metabolic blocks during the model refinement.

3.1.6 Escher

Escher is a web application for the semi-automatic construction and visualization of metabolic pathways [96]. Escher was used to construct a metabolic map of the central carbon metabolism of *A.* sp. T66 to assist in the visualization of the simulated flux distributions.

3.1.7 Gurobi

The Gurobi optimizer is a stand-alone proprietary solver for a wide range of optimization problems, such as linear programming, quadratic programming, and mixed-integer linear programming (MILP). The solver provide interfaces for multiple modeling and programming languages, including MATLAB [97]. It is also one of the supported solvers for a majority of the functions within the COBRA toolbox [70]. During this project, Gurobi version 7.5.2 was used to solve all optimization problems [97].

3.1.8 BLAST

BLAST (Basic Local Alignment Search Tool) is a local alignment algorithm for finding similar sequences in a library of sequences using a query of an amino acid sequence of a protein or the nucleotide sequence of a gene [98]. Although not necessarily awarding an optimal alignment [99], the heuristic algorithm has gathered traction for its speed, enabling rapid identification of similar sequences in ever-growing genomic databases [100]. The algorithm can be divided into three steps: (1) compiling a list of *k*-letter high-scoring words, (2) scanning the database for hits of these words, and (3) extending the resulting hits to identify sequences of high similarity.

In step (1), the algorithm pre-processes the query sequence by removing regions of low compositional complexity, that is, regions of biased overrepresentations of particular amino acids or nucleotides which may award highly significant, but biologically meaningless hits [100]. The query sequence is then fragmented into *k*-letter words and further expanded with a set of ”synonyms” based on possible changes of these initial words due to random mutations [101]. These ”synonyms” are evaluated and discarded if the resulting score based on the substitution matrix (e.g. BLOSUM62 [102]) falls below a given threshold.

In step (2) and (3), the database is scanned for exact matches of these high-scoring matching words. These matches are then extended in both directions until the optimal accumulated score begins to decrease. Hits are then kept or discarded, depending on if their associated score falls below an empirical threshold. The statistical significance of the resulting list of hits are then determined by finding the probability that two random sequences of lengths equal to the query sequence and the entire database, respectively, could generate the calculated scores given the same composition [100, 101]. The database sequences associated with the hits of highest significance are then used in a second local alignment with the query sequence using the Smith-Waterman algorithm. The resulting alignment are given as an output, along with the statistical scores which may be used to evaluate the similarities between the query sequence and the similar sequences in the database [100].

Several BLAST programs exist which varies based on whether the query and database sequences are made up of amino acids or nucleotides [100]. During this project, two of these programs were employed using the online interface available at
https://blast.ncbi.nlm.nih.gov/Blast.cgi [103]; BLASTp and tBLASTn. BLASTp is used to compare an amino acid query sequence against a protein sequence database, while tBLASTn is used to compare an amino acid query sequence against a nucleotide sequence database which is translated in all six reading frames [100]. While BLASTp was used to find similar sequences of the predicted peptides of *A.* sp. T66, tBLASTn was used to identify ”hidden” protein-encoding genes from the genomic sequence, which due to insufficient coverage of the predicted protein-coding sequences were not categorized as unique open reading frames (ORFs).

3.1.9 HECTAR

HECTAR is a statistical tool for predicting the subcellular targeting of proteins from the eukaryotic clade of heterokonts [104]. It proposes a divide-and-conquer approach in which

the assignment is divided into multiple hierarchical layers consisting of already existing prediction methods. The outputs of these layers are then combined using a Support Vector Machine (SVM), awarding a unique prediction into one of five classes of subcellular targeting: signal peptides (secretory pathway), type II signal anchors (membrane anchoring), chloroplast transit peptides, mitochondrion transit peptides and proteins containing no N-terminal signal sequence. Although primarily developed to handle the difficulty of predicting chloroplastic proteins due to the unique bipartite chloroplast target peptide of phototrophic heterokonts [105], the inherent application of multiple prediction tools could presumably strengthen the resulting predictions of proteins originating from heterotrophic heterokonts such as *A.* sp. T66. HECTAR was therefore used to infer the putative localization of the various proteins of the model [104].

3.1.10 DeepLoc

DeepLoc is a prediction algorithm for the subcellular localization of eukaryotic proteins [106]. The basis of the algorithm is a recurrent neural network that takes into account the entire amino acid sequence of the protein, identifying regions of the sequence which are particularly decisive in predicting the subcellular localization. The prediction tool is able to predict ten unique localizations: nucleus, cytoplasm, extracellular, mitochondrion, cell membrane, endoplasmic reticulum, plastid, Golgi apparatus, lysosome/vacuole and peroxisome. DeepLoc was used in combination with HECTAR in the prediction of the subcellular localization of all model protein [106].

3.2 Draft model reconstruction and refinement

3.2.1 Initial draft reconstruction

To obtain an initial draft reconstruction, an already published metabolic model of the closely related *S. limacinum* SR21, iCY1170_DHA [20], was utilized as a template model using method (1) of the RAVEN Toolbox. The peptide sequences of *S. limacinum* was obtained from the JGI Database [107], and together with the predicted peptide sequences of *A.* sp. T66, a reciprocal BLASTp was performed using the functionality of RAVEN with default parameter settings [53]. The resulting blast structure was subsequently used to infer putative homologs by bidirectional best hits between the two organisms using the following parameters: alignment length of at least 200, a minimal E value of 10^{-30}, and a minimum identity of 40%. With the help of the corresponding boolean gene-reaction rules, all reactions from the template model associated with the homologous genes were extracted to form the initial draft model.

3.2.2 Ensuring biomass production

In order for a metabolic model to be able to predict specific growth rates, the production of all biomass precursors has to be made possible. The next step of the draft model curation was therefore the identification of gaps responsible for the prevention of biomass production. Due to insufficient knowledge of the specific biomass composition of *A.* sp. T66, the

biomass reaction of the template organism *S. limacinum* was employed. The production of each biomass precursors were tested by iteratively adding a demand reaction for every component individually. A linear program was subsequently made for every biomass component in which the flux through the corresponding demand reaction was maximized, using the minimal medium given in Table 3.1. Any biomass component unable to be produced would subsequently lead to an objective value of 0.

For every biomass precursor incapable of being produced, the anabolic pathways responsible for their production were meticulously studied in order to identify putative gap filling candidates. This procedure was partly manual in which the metabolic surrounding of these pathways were inspected in the pathway databases of both KEGG and MetaCyc [78, 59], and partly automatic in which a brute force strategy was proposed in order to fill the final gaps.

In the manual approach, every KEGG Orthology (KO) associated with a particular pathway map was fetched from the KEGG database by utilizing the KEGG Application Programming Interface (API) [78]. These KOs are annotations of groups of homologous genes encoding enzymes of a particular metabolic functionality. The associated pathway map was subsequently annotated with the subset of these KOs found within the genome annotation of *A.* sp. T66 in order to visualize the presumptive metabolic capabilities. Additionally, visual inspection of the metabolic network using ModelExplorer was performed in order to identify gaps and metabolic inconsistencies. Gaps in the metabolic network were thereby identified and filled by adding appropriate reactions from both KEGG and MetaCyc. Corresponding candidate genes were found by either using the genome annotation directly, or by BLASTp searches of the predicted peptide sequences of *A.* sp. T66. tBLASTn searches of the genomic nucleotide sequence was also performed to circumvent

Table 3.1: Carbon-limited minimal medium used during the model refinement of the draft reconstruction with associated uptake rates (mmol gDW^{-1} h^{-1}). The growth-limiting carbon uptake rate of 1.4 mmol gDW^{-1} h^{-1} was assumed to be equal to that of iCY1170_DHA.

Metabolite	Uptake rate
D-glucose	1.400
Ammonium	10.00
Phosphate	1000
Sulfate	1000
Proton	1000
Water	1000
Oxygen	1000
Calcium	0.010
Boron	0.010
Magnesium	0.010
Silisium	0.010
Copper	0.010
Potassium	0.010
Sodium	0.010
Iron	0.010

any inadequate coverage of the peptide sequences. When no candidate genes were identified but the reaction was necessary to enable biomass production, it was added as a gap filling reaction.

The automatic approach was implemented by formulating a brute force strategy in order to fill the final gaps. As the template model was able to generate all required biomass precursors, the set of reactions not added to the draft model during the initial stages contains the needed reactions to enable biomass production. Finding a minimal set of these that are needed for biomass production would therefore, upon addition to the reconstruction, allow the model to grow. The method was carried out as follows: (1) the set of reactions from the template model not included during the initial draft reconstruction were added to the model, (2) random subsets from this set of reactions were iteratively removed until the biomass production was prevented, (3) steps (1) and (2) was simulated 1 000 times in order to determine the minimal set of necessary reactions.

3.2.3 Addition of novel metabolic capabilities and gene re-annotation

Obtaining a draft reconstruction from KEGG

To increase the predictive capabilities of the newly reconstructed GSM, comprehensive investigations of additional metabolic functionalities were initiated. To facilitate this process and aid in the identification of candidate reactions and genes, a secondary draft model was constructed using method (2) of the RAVEN Toolbox. This method make use of the KO IDs found in the KEGG database, and attempts to assign these to significantly homologous genes from the organism of interest. The associated reactions are subsequently used to generate a draft model reconstruction [53]. Pre-trained profile hidden Markov models (HMM, Eukaryota, Identity: 100%) generated using multiple alignments of the genes associated with each KO was obtained from [108]. These were later queried using the predicted peptide sequences of *A.* sp. T66. Genes of significant similarity were assigned to the corresponding KOs, and the associated reactions were subsequently used to construct the secondary draft model.

The contents of this model, along with biochemical pathway maps from KEGG and MetaCyc, were used in combination with the genome annotation to identify reactions and genes that should be added to the draft reconstruction. Candidate metabolic reactions were also identified through the inspection of metabolic dead-ends and the associated blocked reactions of the model. Metabolic maps from KEGG was visualized using the appropriate KO IDs from the genome annotation to qualitatively identify those pathways present in *A.* sp. T66. Candidate genes for missing reactions were thereby identified and suitable metabolic reactions were collected, forming a repository of novel metabolic capability to be added to the reconstruction.

Gene re-annotation

To further enhance the accuracy of the gene-reaction associations, as well as assuring quality-control of the newly identified genes and associated reactions, a gene re-annotation was performed. All the metabolic genes present in the draft model, as well those present in the aforementioned list of additional metabolic capabilities, were individually inspected

and evaluated using the following approach:

1. The gene was used as a query for BLASTp searches against the KEGG and UniProt databases using default parameter settings [78, 109]. Hits from KEGG allowed for easy identification of the set of biochemical reactions the encoded enzyme would catalyze, while hits from UniProt contributed with more detailed descriptions in the form of bibliomic data. The resulting best hits were recorded, and the associated biochemical reactions were compared against those already present in the model. Any discrepancies were resolved through the addition or removal of reactions. Genes with inconclusive functionalities due to low or ambiguous hits were removed from the model.

2. Associated Enzyme Comission (EC) numbers were validated, based on the best hits found in UniProt, in addition to those annotated to the KO IDs. These EC numbers were also used as queries against the MetaCyc database, which provided extensive descriptions of the associated enzymes, along with auxiliary enzymatic activities not found in the KEGG database. When appropriate, these additional metabolic capabilities were also added to the model.

3. Subcellular predictions were made for all genes using HECTAR and DeepLoc [104, 106]. In cases of differing predictions, HECTAR took precedence over DeepLoc as it is specifically tailored for predicting the subcellular targeting of proteins from the eukaryotic clade of heterokonts in which the thraustochytrids belong [12, 104]. Qualitative evaluations were also performed in cases of conflicting predictions, or when the subcellular localization could be established from high-quality biochemical information. When the subcellular localization of existing genes in the model was changed, appropriate transport reactions were added if subsections of the associated metabolic pathways now appeared in different compartments.

4. The existence of auxiliary subunits of heteromeric enzyme complexes were inferred based on database and literature surveys of the various enzymes. Putative subunits were subsequently identified, and the associated gene rules were updated accordingly.

5. Candidate isozymes were determined by BLASTp searches against the predicted peptide sequences of *A.* sp. T66 using the gene as a query. Steps 1. - 4. were then repeated for the sequences of significant similarity.

In some cases, these additions led to new metabolic dead-ends, causing the process to take on an iterative nature as more and more reactions and genes were added to the reconstruction.

Transporter identification and incorporation

To more accurately represent the transport mechanisms occuring between the various model compartments, a separate round of protein transporter identification was initiated. The motivation behind this was twofold; first and foremost, the transport reactions that

originated from the template model was primarily in the form of reversible uniport reactions. Other modes of transportation such as symport and antiport mechanisms are highly relevant in biochemical systems in which concentration gradients are established and utilized in order to drive the transport of metabolites [110]. Furthermore, antiport reactions are particularly important with regards to mitochondrial transport, as most transport reactions carried out by members of the mitochondrial carrier (MC) superfamily utilizes this transportation scheme [111]. Secondly, although the KEGG database contain entries for a wide range of metabolite transporters, they hardly include any reaction information on their specificity nor information regarding the associated transport mechanism. Consequently, the resulting draft model generated using the KEGG functionality within the RAVEN Toolbox contained no transport reactions.

To begin with, the set of candidate reactions were extracted using the available information found in the genome annotation. Genes annotated by certain gene ontology (GO) terms relating to the process of metabolite transport were identified. These genes were subsequently subjected to the same gene re-annotation procedure as described previously. Explicit care was made when trying to pinpoint the set of substrates a particular transporter would act on. Database and literature surveys on sequences of highest sequence similarity was the primary source of this information. In addition to BLASTp searches of KEGG and UniProt, the Transporter Classification Database (TCDB) was also queried for similar sequences. The TCDB is a manually curated and freely available database on metabolite transporters from all forms of life. Its entries comprise over 10 000 unique transport systems that are further classified into 1000 transporter families [112]. Significant hits found in these databases were subsequently used to infer the most likely set of metabolites a given transporter acts on, and what mode of transport the given transporter employs.

Updating the polyketide synthase pathway

The biosynthesis of the PUFAs in *A.* sp. T66 occurs by the action of the PKS enzymatic complex (Figure 2.8). Here, the acyl chains are covalently attached to the acyl-carrier protein (ACP) domain of the enzyme complex, which directs the moiety to the various catalytic domains during the successive elongation [16]. The PKS pathway from the template model followed a more simplistic mechanism where the associated enzymatic steps were merged into generic reactions involving inaccurate intermediates. Additionally, the acyl moieties were present as acyl-CoA intermediates, thus disregarding the final hydrolysis of the fatty acid from the ACP and subsequent ATP-driven condensation with CoA. A more accurate representation of the PKS pathway was therefore implemented in which the acyl-CoA intermediates were replaced by acyl-ACPs. Every single catalytic step was also added as unique reactions in order to depict the presumed biochemical reaction mechanism occurring *in vivo*. The final hydrolysis of the acyl chain was also implemented in order for the model to predict more realistic energy demands due to the ensuing condensation with CoA.

Addition of the peroxisome

The draft model initially contained only three compartments; cytosol, mitochondria and an extracellular compartment. Because of its central role in fatty acid metabolism [113], a

peroxisomal compartment was added to the model. To resolve which reactions that should be added to this compartment, two strategies were employed. First and foremost, biochemical literature on the peroxisomal metabolism of eukaryotes was studied extensively to get an overview of the various metabolic roles of the subcellular compartment. Associated genes from well studied organisms such as *Homo sapiens*, *S. cerevisiae* and *Arabidopsis thaliana* were then used as queries for BLASTp searches against the predicted peptides of *A.* sp. T66 to identify putative homologs. Secondly, proteins containing peroxisomal target signals were identified by searching for PTS1 and PTS2 signal sequences. These target sequences are universal signals found in many of the proteins targeted to the peroxisome [114, 115], and are recognized by specific receptor proteins which upon binding, initiates the translocation across the peroxisomal membrane [115]. The associated reactions of the resulting candidate genes were then added to this novel compartment in the model. Subsequent debugging in the form of metabolic gap filling later ensued.

3.2.4 Biomass reformulation

During the initial stages of the draft model refinement, the biomass reaction was taken directly from the template model. Although the two organisms most likely have minor differences in their biomass compositions, significant variations in fatty acid composition are quite normal between different thraustochytrid strains [116]. As an accurate description of the fatty acid composition is central to this project, efforts were made to incorporate the organism-specific fatty acid distributions. A restrictive approach was carried out in which the lipid composition was assumed to be the same as for *S. limacinum*, while the acyl chain distribution was obtained from generic fatty acid methyl ester (FAME) analysis of *A.* sp. T66 cells from batch-fermentation experiments. Lacking detailed experimental data on the quantitative levels of specific lipids of particular chain configurations, it was further assumed that all lipid classes would have the same fractional acyl chain composition.

In the general case, the stoichiometric coefficients of the various acyl moieties were updated to take on the values of the normalized fractions as given in Table 3.2. These correspond to the molar ratios of the various acyl chains. Subsequent reformulation of the stoichiometric weightings of the lipid classes in the biomass function was then performed.

Table 3.2: Normalized fractional fatty acid composition of *A.* sp. T66 obtained from batch-fermentation experiments. These values were used to impose a fatty acid distribution on all lipid forms implemented in the GSM.

Fatty acid	Lipid number	Fractional composition
Tetradecanoate	c14:0	0.121
Hexadecanoate	c16:0	0.314
Hexadecenoate	c16:1(n-7)	0.134
Octadecanoate	c18:0	0.017
Octadecenoate	c18:1(n-7)	0.072
Eicosapentaenoate	c20:5(n-3)	0.006
Docosapentaenoate	c22:5(n-6)	0.049
Docosahexaenoate	c22:6(n-3)	0.287

This was due to the template model utilizing a twofold stoichiometric weighting, both in these acyl-consuming/producing reactions and in the biomass reaction.

The biomass reaction was further split up into five distinct reactions, each containing sets of biomass precursors belonging to a particular category. These five categories consisted of DNA, RNA, amino acids, carbohydrates and lipids. For the DNA and RNA groupings, (deoxy)nucleoside triphosphates ((d)NTPs) were used as reactants, simulating the energetic cost of replication and transcription, respectively. Similarly for the amino acid reaction, aminoacyl-tRNAs were used for each of the 20 amino acids, along with additional energy carriers in the form of adenosine 5'-triphosphate (ATP) and guanosine 5'-triphosphate (GTP) required for the process of translation. For the carbohydrate reaction, UDP-D-galactose and GDP-L-galactose were employed. These energetic demands constitute a subset of the cellular requirements for energy, thus forming a minimal estimate of the GAM. The lipid reaction was further updated from the template model through the addition of free fatty acids, as well as the exclusion of mono- and diacylglycerols, whose experimental levels were unobtainable.

3.3 Metabolic network evaluation

3.3.1 Detection and removal of energy-generating cycles

Stoichiometric models of genome-scale metabolism may contain thermodynamically infeasible cycles which are able to charge intracellular energy carriers without any nutrient consumption [117]. These cycles are products of inaccurate reversibilities of the constituent reactions, which often result from a lack of thermodynamic constraints within the modeling framework. Alternatively, subsets of the cycles could be thermodynamically feasible only under particular environmental conditions (i.e. metabolite concentrations), but not simultaneously [117]. The detection and removal of these cycles is therefore

Table 3.3: Set of dissipation reactions used to identify thermodynamically infeasible EGCs. Each dissipation reaction were iteratively added to the model and subsequently optimized when all exchange reactions were constrained to zero. Any non-zero optimal objective value indicated the presence of an ECG.

Energy carrier	Dissipation reaction
ATP	$ATP + H_2O \longrightarrow ADP + Phosphate + H^+$
CTP	$CTP + H_2O \longrightarrow CDP + Phosphate + H^+$
GTP	$GTP + H_2O \longrightarrow GDP + Phosphate + H^+$
UTP	$UTP + H_2O \longrightarrow UDP + Phosphate + H^+$
ITP	$ITP + H_2O \longrightarrow IDP + Phosphate + H^+$
NADH	$NADH \longrightarrow NAD^+ + H^+$
NADPH	$NADPH \longrightarrow NADP^+ + H^+$
FADH_2	$FADH_2 \longrightarrow FAD + 2\,H^+$
Ubiquinol-9	$Ubiquinol-9 \longrightarrow Ubiquinone-9 + 2\,H^+$
Acetyl-CoA	$Acetyl-CoA + H_2O \longrightarrow acetate + CoA + H^+$
L-glutamate	$L-glutamate + H_2O \longrightarrow 2\text{-}oxoglutarate + Ammonium + 2\,H^+$

paramount in order to prevent an overestimation of the predicted growth rates and yields. Of particular relevance are so-called energy-generating cycles (EGCs), a subset of type II extreme pathways [118], in which internal cycles drives the biochemical charging of metabolic energy carriers without any input of energy [117]. These may be regarded as the thermodynamically infeasible counterpart of futile cycles, where energy is dissipated without being used to drive any biochemical process [119, 120].

Although the presence of these EGCs is quite easy to identify, the source of their occurrence is more of a challenge to single out. Generally, the most common strategy is to add an energy-consuming reaction (e.g. ATP hydrolysis). The reaction is then selected as an objective to be maximized in a generic FBA formulation when all exchange reactions are constrained to zero. Any non-zero flux through this reaction will therefore indicate the existence of an EGC [117]. This approach was carried out for the model reconstruction to identify any EGC.

The set of tested energy carriers and associated dissipation reactions are all stated in Table 3.3. These dissipation reactions were iteratively added to the model reconstruction and subsequently selected as the objective function to be optimized in a closed model (i.e. all exchange reactions constrained to zero). In the case of a non-zero objective value, the minimal set of active reactions were determined by minimizing the sum of absolute fluxes to remove any infeasible type III extreme pathways [121]. The associated flux distribution of the remaining active reactions was subsequently inspected. The directionality of these was then compared and updated to those found in the MetaCyc database. If the EGC still remained after these modifications, a final reaction-deletion strategy was proposed in which a minimal set of reactions to be deleted was identified. This set of reactions was selected by performing a qualitative ranking of their genomic evidence, where secondary enzymatic activities inferred from *in vitro* experiments on homologous proteins obtained the lowest score. Deletions of the reactions with the lowest ranks were then performed until the associated EGC was disrupted.

A nearly identical test was performed when only the uptake of the carbon source was constrained to zero. This was to account for alternative EGCs in which extracellular protons are taken up and used to drive the charging of energy carriers. This can often occur with GSMs in which a particular metabolite is taken up via a symport reaction along with a proton, while concurrently being exported through a uniporter [117]. The flux through the set of dissipation reactions was optimized, and the minimal set of active reactions were identified and inspected as described above.

3.3.2 Evaluating the *in silico* model predictions

Comparisons of *in vivo* and *in silico* growth rates

A key part of the network evaluation stage is to compare model predictions with experimental growth data. Uncovered disparities might highlight deficiencies within the model reconstruction, and further guide the modeler in subsequent reformulations and updates to close in on the gap between experimental and predicted phenotypes [18]. To assess the quality of the constructed model, the predicted rate of biomass production was compared to the specific growth rates obtained from experimental efforts. Measured substrate uptake rates were incorporated as lower bounds on the exchange reactions of the corresponding

model metabolites to allow for the comparative analysis. These experimental measurements were performed by Inga Marie Aasen, a collaborator at SINTEF, by request of the author.

Uptake rates were determined by performing two separate growth experiments using two different carbon sources, glucose and glycerol, with ammonium as a nitrogen source in both experiments. Cells were cultivated in 100 mL medium in 500 mL shaking flasks at 28 °C and 250 rpm. Four samples of 10 mL were extracted during the exponential growth phase and subsequently centrifuged at 3200 g for 10 minutes. The supernatant was frozen for quantitative analysis for contents of carbon source and ammonium. The pellet was washed once in 0.9% NaCl, and the resulting supernatant was carefully removed before drying at 105 °C for 16-20 hours to enable dry weight quantification. Glucose and glycerol were quantified by high performance liquid chromatography (HPLC). Samples were centrifuged and filtered through 0.2 μm syringe filters before analyzed using an Aminex HPX-87-H column (BioRad Laboratories) at 45 °C, and refractive index detection (RID-6A, Shimadzu). Five mmol H_2SO_4 was used as mobile phase at 0.6 mL/min. Ammonium quantities was determined using an enzymatic kit according to the instructions (Megazyme K-AMIAR). The quantified levels of the carbon source and ammonium was subsequently used to estimate the specific uptake rates (calculations available in Appendix B).

3.4 Model employment for phenotypic predictions

3.4.1 Assessing gene essentiality by *in silico* gene deletion analysis

Through the employment of the boolean gene-reaction associations in GSMs, one may study the effect of single- or multiple-gene deletions on the metabolic phenotype under various conditions. These predictions do not only offer information on the properties of the metabolic network, but can also be utilized in the context of metabolic engineering, where the effects of genetic perturbations can be assessed to evaluate the viability of proposed gene knockout strategies.

Although traditional gene essentiality studies primarily focuses on the ternary classification of genes as essential, partly essential or non-essential [122], another interesting evaluation is how the gene deletion affects the flux redistribution in the perturbed wild type network. Employing the proposed method by Xu et al. [123], the deletion impact *p* on the metabolic flux redistribution were calculated for all genes of the model. For a set *R* of metabolic reactions, the deletion impact *p* is defined as

$$p = \sum_{j}^{R} (v'_j - v_j)^2, \tag{3.1}$$

where v'_j and v_j is the flux through reaction *j* in the perturbed and wild type network, respectively. The flux distributions of both the wild type and perturbed network were identified by FBA, allowing for the calculations of the associated *p* values.

Traditional gene essentiality simulations were also performed by evaluating the effect on the growth rate of each single-gene deletion. A threshold of 10% of the wild type growth rate was applied, defining essentiality. Three distinct classes were defined; those

where the growth rate fell below the given threshold (essential genes), those with a growth rate above the threshold, but below that of the wild type network (partially essential genes), and those with no adverse impact on the predicted growth rate (non-essential genes). Subsequent subsystem enrichments of these three classes were then evaluated and compared with the results obtained from the *p* score analysis.

To more accurately reflect the essentiality of cofactor biosynthesis and utilization, the biomass reaction was expanded by adding indispensable cofactors, coenzymes, and inorganic ions - see Appendix A for list of metabolites. Cofactors and coenzymes were added at arbitrarily low levels (stoichiometric coefficients of 10^{-6}), not to alter the growth estimations, but rather to verify the ability of the model to synthesize them. Although the magnitude of the resulting flux values in the associated anabolic pathways are too low to be biologically realistic, the emerging flux distributions should more accurately depict what actually occurs *in vivo*, and enable essentiality predictions of the associated genes.

3.4.2 Genetic interventions for increased lipid production

The production of malonyl-CoA and reducing power in the form of NADPH is regarded as limiting steps for increased lipid production in oleaginous microorganisms [16]. Using the reconstructed model, optimization strategies for increased production of these were investigated using the OptKnock algorithm. OptKnock is a computational framework formulated as a bilevel optimization problem which tries to identify single- or multiple reaction knockouts in which the flux through a target reaction of interest is higher compared to that of the wild type network [124]. The optimization problem operates by reorganizing the connectivity of the metabolic network through reaction deletions such that the target reaction becomes coupled with the cellular objective. Optimizing for the same objective will thereby necessitate an increased flux through the target reaction, theoretically leading to increased levels of the target metabolite. For a GSM containing a set N of reactions and a set M of metabolites, the bilevel mixed-integer optimization problem is expressed mathematically as

$$
\begin{aligned}
&\max_{y_j} && v_{chemical} && \\
&\text{subject to} && \max_{v_j} && v_{objective} \\
& && && \text{subject to} && \sum_{i=j}^{N} S_{ij} v_j = 0 && \forall i \in M \\
& && && && v_j = v_j^{exchange} && \forall j \in \xi \\
& && && && v_{objective} \geq v_{objective}^{target} && \\
& && && && v_{atp} \geq v_{atp\_main} && \\
& && && && v_j^{min} \cdot y_j \leq v_j \leq v_j^{max} \cdot y_j && \forall j \in N \\
& && && y_j = \{0,1\} && \forall j \in N \\
& && && \sum_{i=j}^{N} (1 - y_j) \leq K. &&
\end{aligned}
$$

Here, the goal is to maximize the flux through the target reaction $v_{chemical}$ by selecting the set of active reactions through reaction deletions of up to K deletions. The deletions are expressed using the binary variables y_j, given by

$$y_j = \begin{cases} 1, & \text{if reaction flux } v_j \text{ is active} \\ 0, & \text{if reaction flux } v_j \text{ is inactive}, \forall j \in N. \end{cases} \tag{3.2}$$

This outer problem is subjected to the constraints of the inner problem which seek to maximize a cellular objective $v_{objective}$, subject to the given constraints. The inner problem is a classic FBA formulation in which the first constraint describes the stoichiometric connectivity of the metabolic network subject to the assumptions of mass balance and a pseudo steady-state. The second type of constraints provide bounds for all exchange fluxes v_j (e.g. uptake and secretion rates), in the subset $\xi \subseteq N$ of exchange reactions. The third and fourth constraint assures that $v_{objective}$ and v_{atp} (NGAM), is greater than or equal to a predefined minimum. The final constraint of the inner problem ensures that the flux through an inactive (i.e. deleted) reaction is constrained to zero, while the flux through an active reaction v_j is within the range $[v_j^{min}, v_j^{max}]$. These lower and upper bounds are determined by performing an initial FVA on the metabolic model using the constraints of the inner problem. This nested optimization problem can further be reformulated into a single-level MILP, which is implemented as the OptKnock function within the COBRA toolbox [70].

Four independent double reaction knockout strategies were tested. The first target reaction was selected to be ACC1, which generates malonyl-CoA by carobxylating acetyl-CoA.

The second and third target reaction was the PPP reactions glucose-6-phosphate 1-dehydrogenase (G6PD, EC:1.1.1.49) and 6-phosphogluconate dehydrogenase (PGD, EC:1.1.1.44), both regarded as important producers of NADPH.

The final target reaction was the cytosolic ME, hypothesized to be a key producer of NADPH during lipid accumulation [89]. MOMA was also used to assess the initial flux redistribution following the proposed reaction knockout strategies.

3.4.3 Elucidation of transcriptionally regulated enzymes

To aid in the understanding of the modes of regulation fundamental to the metabolic shift from the exponential growth to lipid accumulation, transcriptomic data was integrated with the metabolic model using the approach proposed in Ref. [125]. In this method, condition-specific models are generated by constraining a set of experimental fluxes in each condition, thus altering the region of feasible flux distributions. Random sampling of the resulting flux spaces gives rise to averages and standard deviations for every flux in the model, which subsequently can be used to infer the statistical significance of change between the conditions. By comparing these changes with the corresponding changes in transcript levels of the associated enzymes, one may infer the sets of metabolic enzymes that are controlled by transcriptional, post-transcriptional or metabolic modes of regulation.

Three condition-specific models were generated for three consecutive stages of a fermentation setup of *A.* sp. T66, where genome-wide transcriptomics analysis had been

performed at the corresponding sampling points [126]. These stages were: the exponential growth phase (E), early onset of lipid accumulation (N1), and late lipid accumulation (N2). The description of the various condition-dependent constraints and underlying assumptions are presented in Table 3.4.

Before the random flux sampling, FVA was performed on all three models to identify the set of reactions involved in type III extreme pathways, which will give rise to unrealistic means and standard deviations during the random sampling procedure. Minimal bounds for these reactions were determined by gradually decreasing/increasing their lower and upper bounds until a feasible solution were obtainable for the individual models (i.e. all condition-specific constraints were maintained). Rather than performing a random, uniform sampling of the feasible solution space using a hit and run (HR) approach, the extreme solutions of the feasible flux distributions were sampled using the Convex Basis (CB) approach. The CB approach tends to be more conservative than the HR algorithm as it generates higher standard deviations of the fluxes [125]. This allows for predictions of higher confidence by reducing the likelihood of falsely assigning a given flux as significantly changed.

Using the CB algorithm, the optimal corner-point feasible solutions was sampled by maximizing the flux through random pairs of reactions 2000 times for each of the three conditions. The resulting means and standard deviations for all fluxes were then used to calculate a Z statistic, which quantifies the significance of flux change between the various

Table 3.4: Condition-dependent rates used to constrain the feasible solution space of the three condition-specific models, allowing for random flux sampling and subsequent comparisons of the differential flux distributions.

Condition	Glycerol[a]	Ammonium[a]	Growth[b]	TAG[a]	Comments
E	2.4432	10	0.1289	0	Exponentially growing cells in a glycerol-limited medium. Growth rate constrained at its optimal value.
N1	2.4432	10	0.04	0.067	Early onset of lipid accumulation, experimental growth rate of 0.04 h^{-1} estimated based on CO_2 levels [126]. Remaining carbon flow assumed to be directed towards TAG production, which is constrained at an optimal level as no measured rates of lipid production are available.
N2	2.4432	0	0	0.097	Nitrogen-depleted medium where all carbon flow is assumed to be used for lipid production. Ammonium uptake constrained to zero, while the TAG production is constrained at its optimal level.

[a] Lower bounds of fluxes given in mmol gDW^{-1} h^{-1}.
[b] Specific growth rates given in h^{-1}.

conditions. The *Z* statistic for a given reaction i is given by

$$Z_i^{flux} = \frac{E_2(v_i) - E_1(v_i)}{\sqrt{Var_2(v_i) + Var_1(v_i)}}, \tag{3.3}$$

where $E_2(v_i)$ and $E_1(v_i)$ are the means of flux v_i of two conditions, and $Var_2(v_i)$ and $Var_1(v_i)$ are the corresponding variances. The significance of transcript level change Z_i^{exp} between the conditions was calculated from the associated p values from the transcriptomics data using the inverse error function

$$Z_i^{exp} = \pm inverf(1 - p_i/2). \tag{3.4}$$

These *Z* statistics were subsequently used to calculate the individual probabilities of each of the three modes of regulation. The probability of transcriptional regulation, P_{tri}, is given by the product of the probability of a metabolic flux changing with the probability of the corresponding enzyme having its transcript level change in the same direction

$$P_{tri} = \Phi(Z_i^{flux})\Phi(Z_i^{exp}), \tag{3.5}$$

where $\Phi(Z)$ is the cumulative Gaussian distribution. For decreasing flux or transcript levels, the absolute values of the *Z* scores were applied. Additionally, if the direction of a reversible reaction changed between the two conditions, a P_{tri} value of zero was assigned by default, as this change only can be ascribed to changing metabolite concentrations.

For post-transcriptional regulation, we obtain the probabilities by multiplying the probability that the transcript levels are changing with the probability that the corresponding flux remains unchanged

$$P_{pri} = \mathrm{erf}(Z_i^{exp})(1 - \mathrm{erf}(Z_i^{flux})). \tag{3.6}$$

Here, $\mathrm{erf}(Z)$ is the error function as we want to assess the probability of change in any direction.

Similarly for metabolic regulation, we obtain the probabilities from the product of the probability of a change in flux, but not in transcript levels

$$P_{mri} = \mathrm{erf}(Z_i^{flux})(1 - \mathrm{erf}(Z_i^{exp})). \tag{3.7}$$

For the probabilities of post-transcriptional and metabolic regulation, the absolute values of the *Z* statistics were employed. Table 3.5 summarizes the various combinations of

Table 3.5: Qualitative assignments of the three modes of regulation based on the various combinations of up-regulation (+), down-regulation (-) and no regulation (=) for flux and transcript levels. TR: transcriptional regulation, PR: post-transcriptional regulation, MR: metabolic regulation.

Exp\Flux	+	-	=
+	TR	MR	PR
-	MR	TR	PR
=	MR	MR	

flux and transcript changes with the resulting forms of regulation. A threshold of 0.95 was selected for all probabilities, resulting in nine independent lists of gene-reaction pairs corresponding to these three modes of regulation for the three condition comparisons; N1 v. E, N2 v. E and N2 v. N1.

Chapter 4

Results and discussion

4.1 Reconstruction and refinement of iVS1191

4.1.1 Constructing the initial draft model

The initial reconstruction generated from the template model iCY1170_DHA contained a total of 1340 reactions, 1099 metabolites and 974 associated genes (available in Supplementary Material as initialDraftModel.mat). Based on the integrated gene-reaction rules, around 92% percent of all reactions from iCY1170_DHA associated with a set of genes were incorporated into the draft reconstruction. However, putative reciprocal best hits for a substantial amount of template genes were not identified.

Further inspection of the remaining 196 genes from iCY1170_DHA identified 103 new putative homologs in *A.* sp. T66. These were determined using reciprocal BLASTp of the predicted peptide sequences with less stringent cutoffs for the alignment lengths. This was done so as to prevent the exclusion of genes whose sequence lengths were too small to be added during the initial draft reconstruction. A more appropriate strategy to avoid this issue could have been to apply a dynamic minimal alignment calculated by a scaling function, whose value would be proportional to the sequence lengths of the query genes. Subsequent BLASTp searches using queries of lengths smaller than the original minimal

Table 4.1: Comparison of the model properties of the template model iCY1170_DHA and the resulting draft reconstructions after iterative modifications.

Model	Note	Reactions	Metabolites	Genes
iCY1170_DHA	Template model	1769	1659	1170
Initial draft model	From template	1340	1099	974
Draft model	Additional genes	1423	1573	1107
Draft model	Transport & exchange	1689	1632	1107
Draft model	Gap filling	1828	1680	1145

cutoff would therefore not automatically be discarded during the initial reconstruction process.

A remaining set of 54 template genes had no significant hits against any predicted peptide of *A.* sp. T66. Although this could reflect genuine genomic disparities between the two organisms, the close phylogenetic distance between them rather hinted at insufficient coverage of the predicted peptide sequences of *A.* sp. T66, preventing the identification of genuine homologous sequences. Consequently, tBLASTn searches of the genomic nucleotide sequences using these genes as queries were performed in order to circumvent this issue. In the end, a total of 30 additional candidate homologs were identified in the genomic sequence of *A.* sp. T66. The corresponding reactions of these genes, along with the reactions associated with the aforementioned set of 103 genes, were subsequently merged with the draft reconstruction, resulting in 83 additional metabolic reactions.

Out of 195 transport reactions of iCY1170_DHA, only 118 had an associated gene-reaction rule. As an initial strategy, all of the remaining transport reactions were therefore transferred to the draft reconstruction. This was done to prevent their exclusion from impeding on the ability of the model to produce any of the necessary biomass components. Similarly, all exchange reactions from iCY1170_DHA were also added to the draft reconstruction. The properties of iCY1170_DHA and the resulting draft reconstructions during these iterative modifications are all listed in Table 4.1.

4.1.2 Gap filling for biomass production

Following the initial round of metabolic gap filling, the model was able to generate all the necessary biomass precursors, with a corresponding specific growth rate of 0.0834 h^{-1}. This was a slight increase from that of iCY1170_DHA, which at the same glucose uptake rate of 1.4 mmol gDW^{-1} h^{-1} predicted a specific growth rate of 0.0812 h^{-1}. Although a minor increase, the *in vivo* exponential growth rates of *A.* sp. T66 is in reality found to be considerably higher than that of *S. limacinum* - see Table 4.6 and [20]. This clearly indicated that the metabolic network was lacking necessary biochemical potential, and consequently, that further additions of metabolic reactions was needed.

Although the primary goal was to enable biomass production, other additional reactions was also introduced to more accurately represent the metabolic network of *A.* sp. T66. To illustrate this curational procedure, consider the following example of how the metabolic gaps were identified and filled.

Following the inclusion of the additional genes, as well as the remaining transport and exchange reactions, the metabolic model was unable to generate 9 of the 49 biomass precursors. One of these metabolites was the steroid derivative stigmasterol. The key precursor for stigmasterol is the isoprenoid squalene, which is synthesized from successive condensations of isoprene units derived from the mevalonate pathway [127]. From squalene, a set of 14 consecutive biochemical modifications are required to generate stigmasterol. As seen in Figure 4.1, the model was deficient in 5 intermediary enzymatic steps needed to synthesize stigmasterol.

The genome annotation was then probed for genetic candidates based on associated KO IDs so as to elucidate the putative metabolic environment found in *A.* sp. T66. This was assisted by color coding the KEGG steroid biosynthesis pathway map to highlight the metabolic capabilities of *A.* sp. T66 (Appendix C, Figure 5.1). Based on this initial

approach, candidate enzymes for two out of the five missing reactions were identified directly from the genome annotation. These were the genes T66009048.1 and T66008786.1, putatively encoding the enzymes methylsterol monooxygenase 1 (SMO1, EC:1.14.18.9), and sterol 22-desaturase (CYP710A, EC:1.14.19.41), respectively. Their exclusion during the initial draft reconstruction was a result of iCY1170_DHA not having any associated genes for these reactions, although candidate genes can be identified in the peptide sequences of *S. limacinum* (Aurli1_80852 and Aurli1_74786, respectively).

For the remaining three reactions, sequences associated with the appropriate KO IDs were used as queries for both BLASTp and tBLASTn searches against the peptide and genome sequences of *A.* sp. T66. This resulted in the identification of a final gene, T66005921.1, putatively encoding a cholestenol δ-isomerase (HYD1, EC:5.3.3.5). Although no candidate enzymes were found for the latter two enzymes 24-methylenesterol *C*-methyltransferase (SMT2, EC:2.1.1.143), and squalene monooxygenase (SQLE, EC:1.14.14.17), the associated reactions were added as gap filling reactions to the model so as to enable the biosynthesis of stigmasterol.

A similar approach was carried out for the remaining 8 biomass precursors that the model initially was unable to generate. These were: d-galactose, l-galactose, l-leucine, l-lysine, l-proline, l-tyrosine, cholesterol and phosphatidylglycerol. In total, 139 reactions were added during this initial gap filling procedure, of which 12 reactions were non-spontaneous gap filling reactions with no identifiable candidate genes. Simulated deletions of the latter reactions, along with the collection of incorporated transport reactions, culminated in a minimal set of 15 gap filling reactions necessary to allow the model to synthesize all biomass precursors (Table 4.2). These simulations were performed by single reaction deletions of this set with a predetermined cutoff of 10% of the wild type specific growth rate defining reaction essentiality [128].

The main concern of metabolic gap filling of GSMs is the inclusion of erroneous biochemical reactions which affords the model with additional metabolic capabilities not found *in vivo*, potentially leading to inaccurate phenotypic predictions. Sufficient efforts therefore need to be made to minimize the number of included gap filling reactions, thereby reducing the likelihood of faulty model predictions. Although present as gap filling reactions with no associated genes, the cytoplasmic transport reactions of magnesium (AUR1515), silicon (AUR1516), boron (AUR1517) and copper (AUR1518) are highly biologically significant as the transported elemental compounds are essential biomass components.

The latter mitochondrial transport of 2-oxobutanoate (AUR1704) is known to occur *in vivo* in other eukaryotes [129], where the transporter is hypothesized to be the pyruvate carrier [130]. On the other hand, the three remaining mitochondrial transport reactions (SLI1423, SLI1465 and AUR1724) could potentially pose a problem as no supporting bibliomic evidence was found for their existence in other eukaryotes. Their requirement may likely to be a result of inaccurate predictions of the subcellular localization of the proteins involved in the surrounding metabolic pathways. However, the associated metabolites of these transport reactions are key intermediates in the linear biosynthetic pathways of the branched-chain amino acids valine, leucine and isoleucine [131]. Although the subcellular localization of the surrounding metabolic reactions might be inaccurate, the effects on the predictive capabilities of the model would presumably be insignificant due to these inter-

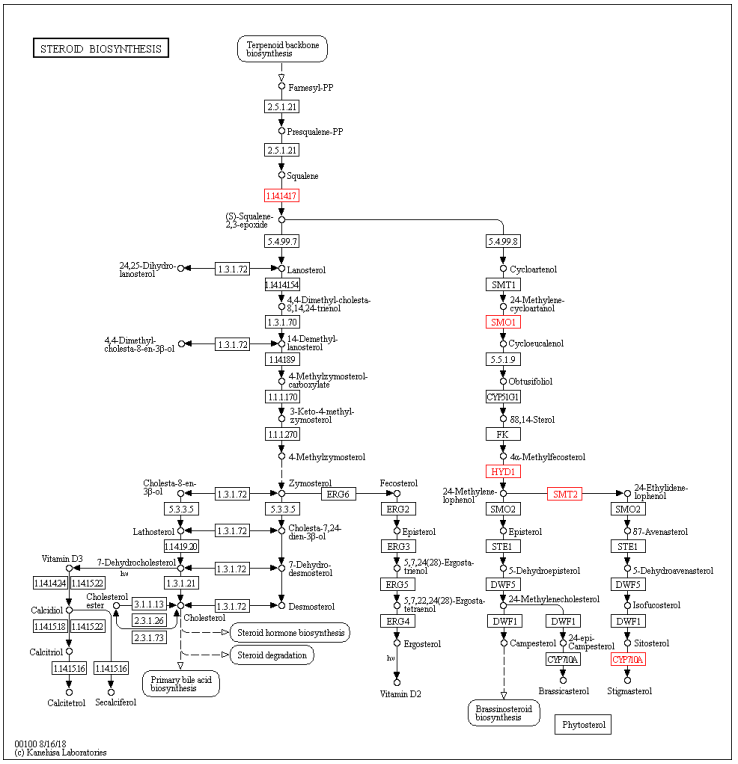


Figure 4.1: Metabolic pathway map of steroid biosynthesis from the KEGG pathway database [78]. Highlighted in red are the five missing enzymatic steps in the draft reconstruction required for the capability of generating stigmasterol; squalene monooxygenase (SQLE, EC:1.14.14.17), methylsterol monooxygenase 1 (SMO1, EC:1.14.18.-), cholestenol δ-isomerase, (HYD1, EC:5.3.3.5), 24-methylenesterol *C*-methyltransferase (SMT2, EC:2.1.1.143), and sterol 22-desaturase (CYP710A, EC:1.14.19.41).

mediary metabolites only taking part in isolated pathways. Similarly, the four gap filling reactions associated with the biosynthesis of steroids (AUR1151, AUR1152, AUR1153 and AUR1164), as well as diaminopimelate dehydrogenase necessary for the biosynthesis of lysine (AUR0392), are all constituents of linear pathways, such that the outcome of their inclusion on the model should only be that of enabling steroid and lysine production, respectively.

The addition of the last two reactions mannose-1-phosphate guanlylytransferase (AUR0068), and GDP-L-galactose phosphorylase (AUR0102), were needed for the model to generate L-galactose. Their inclusion was the result of a final brute force approach in which the minimal set of non-incorporated reactions from iCY1170_DHA needed for biomass production was identified through simulated reaction deletions. Both reactions were present as gap-filling reactions in iCY1170_DHA, suggesting that their responsible enzymes might be elusive in thraustochytrids in general.

Interestingly, the enzyme responsible for mannose-1-phosphate guanlylytransferase

activity has not yet been identified, indicated by the absence of any associated enzyme sequence in KEGG and MetaCyc. Thus, the lack of any genetic evidence may therefore be a result of lacking biochemical knowledge, rather than inadequate coverage of the genomic sequence of *A.* sp. T66. Although one gene, T66010550.1, showed some similiarity with the GDP-L-galactose phosphorylase 2 and 5 of *A. thaliana*, (e values of $8x10^{-6}$ and $1x10^{-4}$, respectively), the similarities were too low and ambiguous to grant the gene a place in the model.

Even though no clear genetic evidence was identified in either the nucleotide nor predicted peptide sequences, the surrounding metabolic proficiency of *A.* sp. T66 suggested that these reactions might be the most realistic candidates to permit the synthesis of L-galactose. Consequently, the reactions were deemed the most likely to occur *in vivo*, and were therefore kept in the model to enable biomass production.

4.1.3 Enhancing the scope of the reconstruction

KEGG model

The secondary draft model was reconstructed using method (2) of the RAVEN Toolbox. Here, reactions annotated to the appropriate KO IDs were fetched based on significant sequence similarities between the predicted peptides of *A.* sp. T66 and the associated sequences in the KEGG database.

The reconstructed model contained a total of 1455 reactions, 1556 metabolites and 1090 genes (available in Supplementary Material as keggDraftModel.mat). Of these genes,

Table 4.2: Minimal set of gap filling reactions of the initial draft model reconstruction necessary for the production of all biomass components. The letters in square brackets indicate between which subcellular compartments the transport reaction occur; c - cytosolic (between the extracellular and cytoplasmic compartment), m - mitochondrial (between the mitochondrial and cytoplasmic compartment).

Model ID	KEGG ID	EC	Reaction name
AUR0068	R00883	2.7.7.22	Mannose-1-phosphate guanylyltransferase (GDP)
AUR0102	R07678	2.7.7.69	GDP-L-galactose phosphorylase
AUR0392	R02755	1.4.1.16	Diaminopimelate dehydrogenase
AUR1151	R02874	1.14.14.17	Squalene monooxygenase
AUR1152	R07495	1.1.1.270	3β-hydroxysteroid 3-dehydrogenase
AUR1153	R07496	-	-
AUR1164	R05776	2.1.1.143	24-methylenesterol *C*-methyltransferase
AUR1704	-	-	(S)-2-isopropylmalate transport [m]
SLI1423	-	-	(R)-2,3-dihydroxy-3-methylbutanoate transport [m]
SLI1465	-	-	(S)-2-aceto-2-hydroxybutanoate transport [m]
AUR1724	-	-	2-oxobutanoate transport [m]
AUR1515	-	-	Magnesium transport [c]
AUR1516	-	-	Silisium transport [c]
AUR1517	-	-	Boron transport [c]
AUR1518	-	-	Copper transport [c]

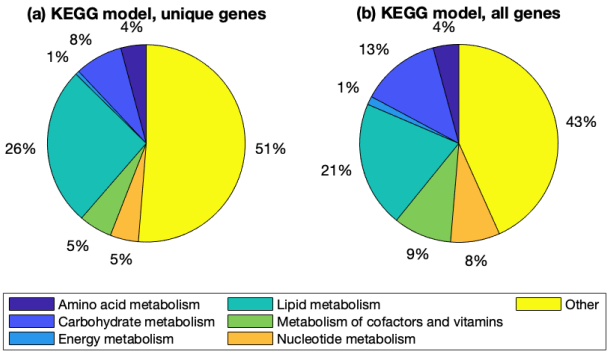


Figure 4.2: Subsystem distributions in the reconstructed draft model of *A.* sp. T66 using the KEGG functionality of the Raven Toolbox. (a) Subsystem distributions of the reactions (n = 626) associated with the genes unique to the KEGG model, (b) subsystem distribution of all reactions (n = 1455) in the KEGG model.

343 were unique to the KEGG model compared to the initial draft reconstruction. This highlights quite a substantial difference in genetic coverage between the two models, and suggests that the metabolic capabilities of the initial reconstruction might be insufficient.

When investigating the subsystem distribution of the reactions annotated to this set of genes, they turned out to be highly enriched with reactions associated with the subsystem 'Other' (Figure 4.3). This generic subsystem consists of several pathways associated with secondary metabolism, which in iCY1170_DHA, and subsequently the draft reconstruction, were predominately absent. As the impact of these peripheral reactions on the flux distribution of the model was insignificant due to them being members of disconnected, and consequently blocked, components of the metabolic network, little effort was presumably made to identify these during the reconstruction of iCY1170_DHA. The rather large subsection of reactions of the lipid metabolism associated with the unique genes in the KEGG model was probably a consequence of a generic specificity of both the biosynthetic and catabolic enzymes, which seemingly hints at the existence of complementary isoenzymes unique to the KEGG model, rather than inadequate reaction coverage of the initial draft reconstruction.

On the same note, the initial draft model contained 398 unique genes compared to the KEGG reconstruction. The subsystem distribution of the associated reactions were somewhat enriched for both lipid and nucleotide metabolism (Figure 4.3). The reason behind the former is most likely the implementation of the PKS pathway for PUFA biosynthesis in *A.* sp. T66, which accounts for 26 independent enzymatic steps which are not found in the KEGG database. Additional reactions associated with the ATP-driven charging of acyl-CoA moieties are also relatively absent in the database for the set of fatty acids implemented in the model. The enrichment of reactions associated with the nucleotide metabolism was due to the existence of non-specific hydrolytic ribonucleoside hydrolases and 5'-nucleotidases, which were annotated to a large set of reactions in the model, often occurring in multiple compartments.

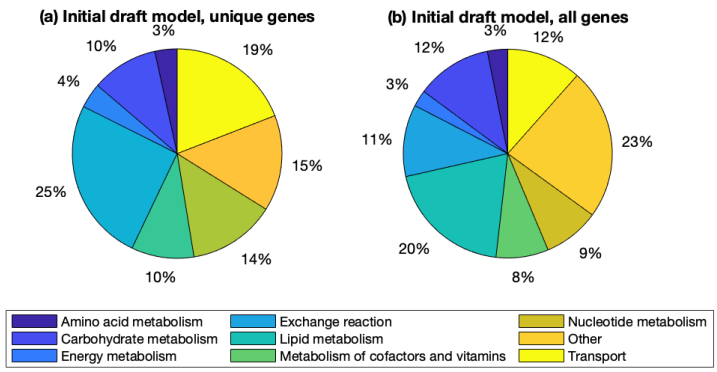


Figure 4.3: Subsystem distributions of the initial draft model reconstructed from iCY1170_DHA. (a) Subsystem distributions of the reactions (n = 466) associated with the genes unique to the initial draft model, (b) subsystem distribution of all reactions (n = 1828) in the initial draft model.

The rather substantial set of unique genes in the two draft reconstructions hints at a shortcoming of only employing one of the two methods in order to generate a draft model of sufficient quality. While the KEGG database consist of a large repository of high-quality biochemical information, it is not specifically tailored to assist in the reconstruction of metabolic models. This is partly reflected in the lack of implemented transport reactions, which plays a key role in the genesis of a genome-scale metabolic networks, especially for the compartmentalized metabolism of eukaryotes. Associated with this is also information on the subcellular localization of proteins, which generally is missing in the KEGG database. Consequently, the largest difference between the two models was the exclusion of any form of transport reactions in the metabolic reconstruction obtained from KEGG, along with no information on the subcellular localization of any of the reactions in the model. Similarly, only utilizing template models for the reconstruction of a draft model may also result in the exclusion of necessary metabolic reactions. Although these models contain additional information not found in biochemical databases such as KEGG, their quality provide an upper bound on the quality of the resulting draft reconstruction. In hindsight, it may therefore have been beneficial to sacrifice the small phylogenetic distance between *A.* sp. T66 and *S. limacinum* for models of higher quality and scope. A more appropriate strategy could have been to employ multiple template models from a taxonomically diverse range of organisms, and in doing so, minimize the number of excluded genes and associated reactions.

Gene re-annotation reveals the importance of manual curation

The motivation behind the manual re-annotation of the metabolic genes of *A.* sp. T66 was threefold. Firstly, during the manual gap filling procedure, as well as throughout the identification of novel metabolic capabilities, the quality of iCY1170_DHA was questioned due to the apparent scarcity of manual curation of the model reactions and associated genes. Several genes seemed to be annotated to erroneous sets of reactions, and the subcellu-

lar localization of a large number of reactions seemed to be based purely on automatic predictions.

A telling example was the subcellular localization of the electron-transferring-flavoprotein dehydrogenase (ETFDH, EC:1.5.5.1), which transfers electrons from multiple mitochondrial flavoprotein dehydrogenases, thus coupling the degradation of fatty acids and certain amino acids with oxidative phosphorylation [132]. In iCY1170_DHA, this reaction was localized to the cytosol, effectively preventing the model from properly degrading these metabolites.

Secondly, during the comparison between the genes of the two draft reconstructions, considerable disparities were found between the genes that were present in both models. This was particularly true for proteins containing multiple catalytic domains, where the KEGG functionality in RAVEN appeared to have difficulties assigning appropriate KO IDs to these multi-functional enzymes.

Lastly, inferring enzymatic activity from the genome annotation of *A.* sp. T66 was in certain cases quite difficult, primarily because of limited information in the annotation, but also due to certain occurrences of inaccurate KO ID assignments and outdated database information.

Although highly laborious, the process of manual re-annotation proved to be a fruitful strategy. In total, high quality annotations on all the model genes were collected, strengthening the qualitative properties of the model, and reinforcing the confidence of the included reactions compared to that of the automatically assembled draft reconstructions. This repository contains information on the candidate metabolic capabilities of all the metabolic genes in the model, in certain cases with associated bibliomic references from sequences of highest similarity. The collection of manual re-annotations can be found in the Supplementary Material as 'Gene Re-annotation.xlsx'.

During the manual re-annotation and the initial rounds of model refinement (Table 4.1), 396 genes not present in the initial draft reconstruction were identified and added to the reconstruction. A substantial set of these were candidate isozymes of already incorporated genes, which presumably was lacking from the initial draft reconstruction as the strategy of bidirectional best hits only identifies sequences of highest similarity. 100 of these genes originated from the KEGG model, whose criteria for inclusion is significant sequence similarity alone, not reciprocality. Consequently, the identification of isoenzymes is therefore greatly facilitated by the KEGG functionality, and subsequently better than the other method which necessitates a reciprocal counter-part in either of the given template models in order for an isoenzyme to be identified and incorporated. This demonstrates the complementarity of the two methods, clearly suggesting that the preferred strategy for automatic reconstruction of GSMs is the concurrent employment of them both.

The subsystem distribution of the associated reactions of these novel genes were highly enriched in transport (27%), as well as lipid metabolism (28%) (Figure 4.4). The former was the result of extensive elucidation and incorporation of metabolite transporters, which in iCY1170_DHA was rather crudely implemented, and furthermore, completely lacking in the model obtained from KEGG. The latter was primarily due to the updated implementation of the PKS pathway with accurate metabolic intermediates and surrounding reaction steps, the addition of the mitochondrial fatty acid biosynthetic pathway, as well as the inclusion of the peroxisome which harbours the peroxisomal β-oxidation of long-chain fatty

acids [113].

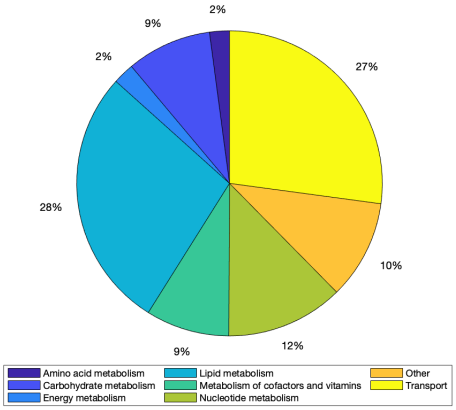


Figure 4.4: Subsystem distribution of the reactions (n = 667) associated with the 396 novel genes identified during the manual gene re-annotation.

In addition to the identification of new genes, 179 genes from the initial draft reconstruction were removed based on erroneous or inconclusive functionalities. Similarly, 345 genes from the KEGG model were not included in the final reconstruction. However, their exclusion were in large extent a consequence of them being of a generic nature, and sometimes involving polymers or protein modifications, which in genome-scale modeling terms is highly irrelevant. Certain cases of erroneous KO ID assignments did also occur in the KEGG model reconstruction, as elucidated through the gene re-annotation procedure. This might be a consequence of using an outdated pre-trained profile HMM, or that inaccurate KO ID assignments were caused by only considering eukaryotic sequences. These numbers corresponds to a rather substantial amount of the genes in these two draft reconstructions (18.4% and 31.6%, respectively), and further illustrate the challenges associated with solely relying on automatic reconstruction of GSMs.

In certain cases, the associated reactions of a given gene were accurate according to the available biochemical information. However, the gene rules were sometimes incorrect, often as a result of the exclusion of auxiliary subunits of both regulatory and catalytic activity. Their identification allowed for more realistic gene-reaction associations, and will greatly increase the validity of the *in silico* analysis of gene essentiality.

As an example, consider the mitochondrial enzyme acetolactate synthase (ALS, EC:2.2.1.6) which catalyze the first step of the biosynthesis of the branched-chain amino acids. This enzyme consist of a larger catalytic subunit, and a smaller regulatory subunit [133]. In *A.* sp. T66, three genes encoding the catalytic subunit of ALS were identified; T66001403.1, T66001404.1 and T66005295.1, along with one gene encoding the regulatory subunit, T66000489.1. Based on iCY1170_DHA, the associated gene rule were constructed exclusively using AND logic

T66001403.1 and T66001404.1 and T66005295.1 and T66000489.1

Using the obtained biological information on the heterodimeric enzymatic complex, the gene rule was updated accordingly

(T66001403.1 or T66001404.1 or T66005295.1) and T66000489.1

The inspected sets of reactions were also examined to ensure mass balance, and updated by either adding additional metabolites (e.g. protons, water), or altering the existing stoichiometries. As a general rule, both the stoichiometry of the reactions and the chemical formula of each metabolite was based on those found in the MetaCyc database. These metabolites are protonated at a reference pH level of 7.3, representing that of the average cytoplasmic compartment [59]. While the subcellular pH levels of the other compartments implemented in the model will be different from this reference state, it was assumed to be constant because of the lack of any experimental data explicitely stating otherwise.

While the detailed reaction mechanisms of the majority of biochemical reactions have been established, accurate stoichiometries of certain reactions still remains undiscovered. Although these incomplete reactions were also present in this reconstructed GSM, the potential impact of the calculated flux distributions is however limited, as these tended to be part of peripheral and isolated pathways, which subsequently were unable to carry a non-zero flux. Certain reactions capable of carrying flux were still present as incomplete reactions in the model (e.g. methylsterol monooxygenase, AUR1141 and AUR1145), which should be updated when appropriate biochemical information on their metabolic transformations is available.

Properties of the peroxisomal compartment

Following extensive literature reviews and identification of proteins that were hypothesized to be targeted to the peroxisome, a peroxisomal compartment was added to the model reconstruction. This compartment harboured 135 unique reactions, consisting of 92 internal and 43 transport reactions, respectively. The metabolic capabilities of the compartment was mainly associated with that of β-oxidation of fatty acids, along with the

Table 4.3: Qualitative assessment of growth on various carbon sources that enters the central carbon metabolism at the level of the two-carbon metabolite acetyl-CoA. Uptake rates were arbitrarily set to 1 mmol gDW^{-1} h^{-1}. Growth and non-growth is denoted by + and -, respectively.

Carbon source	Draft reconstruction[a]	iVS1191	iCY1170_DHA
Tetradecanoate (c14:0)	-	+	-
Hexadecanoate (c16:0)	-	+	-
Hexadecenoate (c16:1(n-7))	-	+	-
Octadecanoate (c18:0)	+	+	-
Octadecenoate (c18:1(n-7))	-	+	-
Eicosapentaenoate (c20:5(n-3))	-	+	-
Docosapentaenoate (c22:5(n-6))	-	+	-
Docosahexaenoate (c22:6(n-3))	-	+	-
Acetate	+	+	+

[a] Draft reconstruction following the initial gap filling to enable biomass production.

associated glyxoylate shunt, which allowed for the biosynthesis of carbohydrates from the generated acetyl-entities of the β-oxidation pathway, along with medium containing other two-carbon compounds such as acetate. The constituent reactions of the latter anaplerotic cycle were predicted to be localized to the peroxisome, as well as the cytosol and mitochondria, suggesting a rather complex interexchange of the metabolic intermediates between these compartments.

While the reactions associated with the glyoxylate cycle were present in the template reconstruction, allowing for growth on two-carbon compounds such as acetate, the model was unable to grow on any of the implemented fatty acids due to insufficient incorporation of necessary transport reactions, preventing the mitochondrial degradation of these acyl chains. Additionally, indispensable enzymatic steps required for the oxidation of unsaturated fatty acids such as Δ^3-Δ^2-enoyl-CoA isomerase (ECI1/ECI2, EC:5.3.3.8) and 2,4-dienoyl-CoA reductase (DECR1/DECR2, EC:1.3.1.34) were also not implemented, preventing the metabolic network from being able to degrade the collection of fatty acids that it produces. When the required reactions were implemented along with the peroxisomal compartment, the reconstructed model was able to grow on all implemented fatty acids (see Table 4.3).

Based on the available genomic data, the β-oxidation of fatty acids in *A.* sp. T66 is situated in both the peroxisomal and mitochondrial compartment. However, the chain length specificities of these two enzymatic machineries remains to be elucidated. A decision was subsequently made, similar to that of the GSM of *Phaeodactylum tricornutum*, iLB1027_lipid [134], in which acyl chains of length 20 and longer were assumed to initially be degraded solely in the peroxisomal compartment. The presumed end product octanoyl-CoA is then exported out of the peroxisome for further degradation in the mitochondrial β-oxidation pathway, due to the presence of a peroxisomal carnitine O-octanoyltransferase (CROT, EC:2.3.1.137). The genuine substrate specificities of the two β-oxidation pathways are most likely overlapping. However, barring any genuine biochemical evidence of their acyl-chain preferences, this differentiation was deemed the most likely based on the available genomic information.

Resolving the biosynthetic pathway of PUFAs

The updated PKS pathway contained a total of 81 unique metabolic reactions, with 75 associated acyl-ACP intermediates. To distinguish the nine initial enzymatic steps which are similar to that of the FAS complex, a duplicated set of metabolites and reactions were

Table 4.4: Candidate genes encoding the enzymatic subunits of the PKS complex responsible for the biosynthesis of the PUFAs in *A.* sp. T66.

PKS subunit	Gene	Homologous gene[a]	Sequence identity (%)	Citation
A	T66011701	KX651612.1	99.96	[135]
B	T66005413.1	KX651613.1	100	[135]
C	T66011702	KX651614.1	100	[135]
acpT	T66011703	KX651615.1	99.31	[135]

[a] GenBank accession number.

added to the reconstruction. The biosynthetic pathway was able to synthesize all of the PUFAs that *A.* sp. T66 produces during lipid accumulation [90]. These consists of the the two ω-3 fatty acids eicosapentaenoate (c20:5(n-3)) and docosahexaenoate (c22:6(n-3)), and the ω-6 fatty acid docosapentaenoate (c22:5(n-6)). The resulting pathway thus diverges early on in the biosynthetic process by either retaining the unsaturated π-bond at the n-3 position, generating eicosapentaenoate and docosahexaenoate, or reducing it, eventually producing docosapentaenoate (Appendix D, Figure 5.2).

In iCY1170_DHA, 15 separate genes were present in the associated gene rules of the PKS pathway. However, upon further inspection it was revealed that many of these enzymes contained rather generic PKS domains, which did not seem to constitute either of the three catalytic subunits of the PKS complex. Through the process of manual gene re-annotation, three candidate genes were found in the genome of *A.* sp. T66 which showed significant sequence similarity with the PUFA PKS subunits of *Thraustochytrium* sp. 26185 (Table 4.4) [135]. Along with an additional gene encoding the auxilliary phosphopantetheinyl transferase (acpT) subunit, which was not present in iCY1170_DHA, these four genes were used to replace the previous set of 15.

4.1.4 Eliminating erroneous EGCs

Upon thorough inspection, it was revealed that the reconstructed model was able to charge all the energy carriers in Table 3.3 without any input of reduced carbon. The responsible EGCs contained a minimal set of 280 - 408 reactions, depending on the particular energy carrier, emphasizing the complexity of the emergent EGCs of stoichiometric models of metabolism. These thermodynamically infeasible cycles concertedly operated by taking up extracellular protons by means of a flux coupling between a proton-coupled transport reaction of a given metabolite, and a reversible uniport reaction of the same compound. The set of reactions then utilized this flow of protons to drive a continuous generation of cytosolic reducing power in the form of NAD(P)H. The retained free energy in these were subsequently used to drive the biochemical charging of the other energy carriers. In the example of ATP generation, the associated electrons were shuttled into the mitochondria via the available substrate-shuttles (e.g. glycerol-3-phosphate shuttle), being used to force the export of protons into the intermembrane space in the respiratory chain, thus driving the ensuing charging of ATP. Due to the presence of these EGCs, the predicted growth rate of the metabolic model increased by around 34%, clearly indicating the importance of its removal for accurate predictions of metabolic phenotypes.

The flux-carrying reactions were inspected by finding the set of reactions present in all of the emergent EGCs, which culminated in a list of 96 unique reactions. From these, a qualitative ranking of the reactions and associated genes were performed, based on an evaluation of the likelihood that the encoded enzymes in fact catalyze their associated reactions *in vivo*. From the identified list, one reaction was singled out as a prime candidate for removal. This reaction was an auxiliary enzymatic activity of the cytosolic enzyme serine hydroxymethyltransferase (SHM1, EC:2.1.2.1), where it converts 5,10-methenyl-tetrahydrofolate to 5-formyl-tetrahydrofolate in the presence of glycine. This catalytic activity has been found in the homologous enzyme of both rabbit and *E. coli* [136, 137], and was therefore assumed to also occur in *A.* sp. T66. This secondary activity is however merely mentioned as a sidenote in MetaCyc, not being explicitly defined as a catalytic

activity of the enzyme in the database. The reaction was consequently deemed to be that with which most uncertainty resided, and was subsequently selected for elimination.

Following its removal, all previously present EGCs were disrupted, preventing the infeasible charging of the aforementioned energy carriers. This approach should however only be considered a temporary solution to the issue of EGCs in iVS1191. More accurate thermodynamic data based on measured metabolite concentrations should, when available, be included in future refinements of the model to alter the reaction directionalities so as to allow for the inclusion of this reaction if it turns out to be a genuine activity of the enzyme, without the occurrence of these infeasible cycles.

To counteract the possibility of obtaining additional EGCs, a mitochondrial intermembrane compartment were added for the protons used to drive the charging of the energy carrier ATP during oxidative phosphorylation. This is a similar approach to that used in human GSM RECON3D [138], where the transfer of protons from the mitochondrial matrix exports the hydrogen ions into this separate compartment, rather than to the cytosol. Import of extracellular protons into the cytosol due to coupled transport reactions can subsequently not be used directly to drive the electrochemical charging of ATP, thus reducing the likelihood of thermodynamically infeasible EGCs following future refinements of the model.

4.2 Properties of the final model reconstruction

4.2.1 Major advancements in both coverage and scope

The final model reconstruction iVS1191 contained a total of 1668 metabolites, 2093 reactions (1455 metabolic reactions, 416 transport reactions and 232 exchange reactions), and 1191 associated genes (available in Supplementary Material as 'iVS1191.xml'). This highlights a significant increase in coverage compared to iCY1170_DHA (Table 4.5). Most noticeable is the additional number of reactions. All three categories (i.e. metabolic, transport and exchange) have expanded their number of reactions, where the number of transport reactions shows the largest increase.

The modest increase in the amount of genes can be explained by the process of manual curation. Here, a large set of genes with sub par hits against enzymes of known functionalities was removed during the process. At the same time, several genes were added, either to replace those that were removed, or as associated genes of novel metabolic reactions not present in iCY1170_DHA. Consequently, although the difference might seem insignificant, considerable alterations have been performed so as to more accurately reflect genuine biological organization with these gene-reaction associations.

To assess the qualitative difference between the two models, the distribution of the 9 main subsystem classes was compared. From the plot in Figure 4.5, one may notice minor alterations in the subsystem distributions of the two models, clearly indicating the close phylogenetic relationship between the two organisms. However, iVS1191 shows a significant increase in the number of transport reactions, while the remaining subsystems show a slight decrease. The reason behind this is twofold. Firstly, the inclusion of a peroxisomal compartment added 43 transport reactions to the model. Secondly, and most importantly, the extensive curation of the inter-exchange of metabolites between the remaining com-

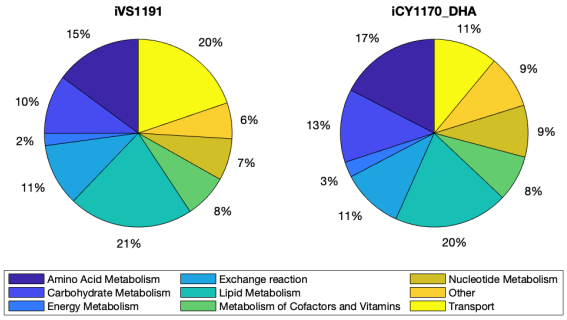


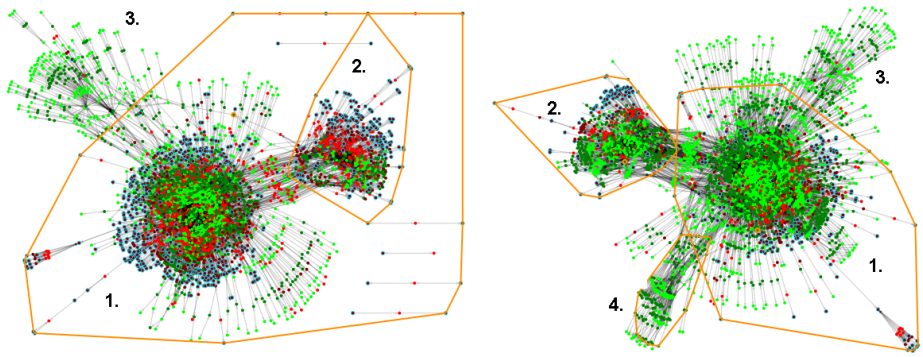
Figure 4.5: Subsystem distribution of the 9 main classes of subsystems from the KEGG PATHWAY database in iVS1191 and iCY1170_DHA. The subsystem 'Other' contains Glycan Biosynthesis and Metabolism, Metabolism of Terpenoids and Polyketides, Biosynthesis of other secondary metabolites, Metabolism of Other Amino Acids, Xenobiotics Biodegradation and Metabolism, and Metabolism of terpenoids and polyketides.

partments during the manual gap filling resulted in a net addition of 175 cytoplasmic and mitochondrial transport reactions. Although the number of transport reactions more than doubled, the fraction of these transport reactions with an associated gene rule increased significantly.

Of the cytoplasmic transport reactions, about 75% has a corresponding gene rule in iVS1191, compared to around 65% for iCY1170_DHA. The fraction only increased slightly from about 55% to 57% for the mitochondrial transport reactions. This limited increase can mainly be attributed to the difficulty of assigning appropriate substrates for a large subset of the transporters belonging to the MC family, due to high degrees of sequence homology between the candidate proteins [139]. Furthermore, as candidate transporters for several metabolites that are known to cross the mitochondrial inner membrane still remain to be elucidated [140], one would expect that a significant proportion of the necessary transport reactions will lack gene associations.

While the ORF coverage of iVS1191 did show a pronounced increase from that of iCY1170_DHA (Table 4.5), this is most likely an effect of a smaller genome, and a corresponding reduction in the number of predicted ORFs. Whereas the genome size of *S. limacinum* is 60.93 Mb (mega base pairs), with 14,859 unique ORFs [107], the genome size of *A.* sp. T66 is merely 43 Mb, with a corresponding list of only 11,683 of predicted ORFs [141]. Although this could imply a deficient genome coverage, the ORF coverage is comparable to other high-quality GSMs of the oleaginous species *Mortierella alpina* and *Yarrowia lipolytica* [142, 135], each possessing an ORF coverage of 9.5% and 9.6%, respectively [20].

A considerable reduction in the number of blocked reactions and associated dead-end metabolites was accomplished during the model reconstruction and curation. In iCY1170_DHA, a substantial subsection of the reactions were found to be blocked, constituting around 44% of all model reactions. Of the metabolites, around 38% were categorized as metabolic dead-ends. Even with a substantial increase in the number of reactions in iVS1191, only around 16% were found to be blocked. Similarly, merely 17% of the



(a) iCY1170_DHA **(b)** iVS1191

Figure 4.6: Visualization of the metabolic reconstructions in ModelExplorer of (a) iCY1170_DHA, and (b) iVS1191. Both reconstructions were converted to open models in which all exchange reactions were set to be unconstrained. Blocked reactions are indicated as red circles, while those that are able to carry flux are indicated in light green. The metabolites are color coded in a similar fashion: blocked = dark red, non-blocked = dark green, where the dead-end metabolites also are highlighted by a surrounding blue edge color. The compartments of the models are indicated by the yellow boarders: (1.) cytosol, (2.) mitochondria, (3.) extracellular, and (4.) peroxisome. The mitochondrial intermembrane compartment of iVS1191 is not visible.

metabolites were found to be metabolic dead-ends.

To illustrate this drastic reduction, the two reconstructions were visualized in Model-

Table 4.5: Comparison of the features of the final model reconstruction iVS1191 and iCY1170_DHA. Dead-end metabolites are metabolites that are unable to be consumed or produced as they are constituents of only one model reaction. The blocked reactions were identified by having lower and upper flux bounds of zero when running FVA on an open model (i.e. all exchange reactions left unconstrained).

Features	iVS1191	iCY1170_DHA
Compartments	5	3
Genes	1191	1170
Metabolites	1668	1659
Reactions	2093	1769
Metabolic	1445	1386
Transport	416	195
Exchange	232	188
Blocked reactions	339	769
Dead-end metabolites	280	628
ORF coverage (%)[a]	10.2	7.9

[a] Gene sequences found in the genome sequence, but not as unique ORFs, were also included to account for insufficient coverage.

Explorer when all their exchange fluxes were unconstrained (Figure 4.6). Here, one may notice how larger parts of the metabolic network of *A.* sp. T66 now are able to carry flux, compared to that of iCY1170_DHA in which considerable sections of the metabolic network are blocked. While the mitochondrial compartment in iCY1170_DHA in large extent is metabolically dead, the reactions of this compartment in the iVS1191 is to a much greater extent able to take on non-zero flux values. These changes are prime examples of how the newly constructed model is able to utilize a greater subsets of its metabolic capability, expectedly giving rise to more accurate phenotypic predictions.

Although there are considerable advancements in the network connectivity because of extensive metabolic gap filling and manual curation, significant proportions of the metabolic network of iVS1191 are still unable to carry flux (around 16% of all reactions). For example, there is a fairly substantial proportion of blocked reactions and associated dead-end metabolites in the mitochondrial compartment of iVS1191, visible as a cluster on top of the compartment in Figure 4.6 **(b)**. This is indicative of a metabolic model that is still in need of future refinements when more bibliomic and biochemical information is available for the organism.

4.2.2 Growth rate predictions suggest insufficient maintenance energies, and hints at an unrealistic biomass composition

Incorporation of the experimentally determined specific uptake rates allowed for a comparative analysis of the *in silico* growth rate predictions with the measured specific growth rates (see Appendix B for calculations). While the ratios between the measured and predicted growth rates on the two carbon sources were very similar (1.09 versus 1.04), the predicted rates of biomass production overestimated the measured rates by about 7% and 14%, respectively (Table 4.6). A reasonable explanation for these exaggerated predictions can partly be explained by underestimating the energy requirements of the model. For example, the model does not include a non-zero lower bound for ATP hydrolysis to simulate the NGAM demands, as no available chemostat data exist for *A.* sp. T66. Similarly, the estimated GAM of the biomass reaction is only based on the implicit ATP requirements to enable synthesis of the cellular macromolecules. Based on the biomass formulation in the model, this GAM estimate constitute only 7.35 mmol ATP gDW^{-1} h^{-1}. As a comparison, the GAM utilized in the GSM of the heterokont *P. tricornutum* were 29.89 mmol ATP gDW^{-1} h^{-1}, with an additional NGAM of 1.5 mmol ATP gDW^{-1} h^{-1} [143]. Assuming comparable maintenance demands for *A.* sp. T66, these values were incorporated into the model to evaluate the impact on the predicted growth rates. As seen in Table 4.6, this resulted in a drastic reduction in growth rate compared to that of the initial predictions.

While the rather naive incorporation of the maintenance requirements likely constitute an overestimation, the severe decline in biomass production could also be ascribed to a dubious biomass composition. In iVS1191, lipids constitute an entire 43% of the biomass. However, for *A.* sp. T66, the lipid content of exponentially growing cells is merely 13%, not reaching these levels until the late lipid accumulation phase [90]. Similarly, the large contents of carbohydrates (32% of dry cell weight) suggest a biomass composition deprived of nitrogen, probably being closer to that of late lipid accumulation, rather than exponentially growing cells. The protein content is also expected to be larger than the

current levels of around 12%, which would increase the implicit GAM of the biomass reaction, closing in on the gap between the predicted and measured growth rates.

The experimental measurements also allude to a similar conclusion as the specific uptake rate of the nitrogen source ammonium was measured to be around 1.11 mmol gDW^{-1} h^{-1} and 0.99 mmol gDW^{-1} h^{-1} on glucose and glycerol, respectively (Appendix B). However, the corresponding simulated exchange rates in iVS1191 was only 0.392 mmol gDW^{-1} h^{-1} and 0.378 mmol gDW^{-1} h^{-1} at the optimal growth rates. This is indicative of a biomass composition deprived of nitrogenous compounds, and suggests that the implemented biomass is most likely erroneous.

4.2.3 Gene essentiality analysis reveals metabolic robustness and adaptability

Simulated gene essentiality on the minimal glucose medium defined in Appendix E, Table 5.3, identified 768 unique genes as non-essential, 230 as partially essential, and 193 as essential. The subsystem distribution of these categorical classes can be seen in Figure 4.7. For the majority of the subsystems, there was a significant enrichment of non-essential genes, suggesting that the metabolic network of *A.* sp. T66 in many cases are rather impervious to genetic perturbations. Additionally, the high number of partially essential genes, particularly in the subsystems 'Carbohydrate Metabolism' and 'Energy Metabolism', suggest that the network is fairly adaptable by being able to redistribute its metabolic fluxes to accommodate the genetic disruptions.

Of the essential genes, the largest subset is associated with amino acid metabolism. Although this partly can be explained by the sheer number of genes associated with these metabolic pathways (300 unique genes), a more satisfactory explanation is a general scarcity of complementary isoenzymes, as well as a predominance of linear pathways with limited metabolic flexibility. In fact, of the 294 unique reactions of the amino acid metabolism associated with a set of genes, only 21% possesses complementary isozymes, while for the remaining reactions of the metabolic network, this value is around 40%.

A similar case can be argued for the genes of the subsystem 'Metabolism of Cofactors and Vitamins', where a total of 48 out of 139 are defined as essential. Although the biological essentiality of the biosynthetic enzymes of cofactors and vitamins is highly evolutionary conserved [144], the enrichment of amino acid metabolism is rather unconventional. However, considering that dead organic matter is one of the main ecological habitats of the thraustrochytrids [145], a reasonable explanation might be limited selection pressure for increased flexibility of amino acid biosynthesis due to a sustained availability

Table 4.6: Comparison of experimental and *in silico* specific growth rates (h^{-1}) on various minimal media, using measured uptake fluxes to constrain the model (mmol gDW^{-1} h^{-1}).

Medium	Uptake flux	Experimental	*In silico*	*In silico*[a]
Minimal glucose	1.44	0.124	0.134	0.016
Minimal glycerol	2.44	0.114	0.129	0.015

[a] After the incorporation of additional growth- and non-growth associated maintenance demands.

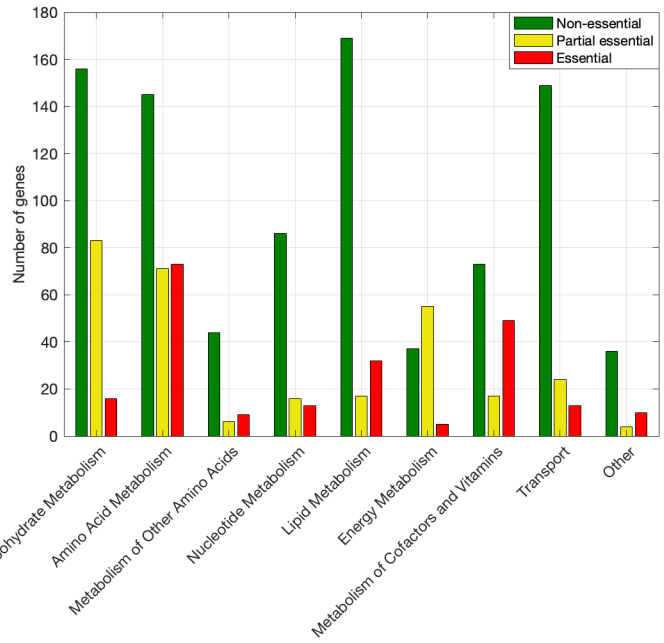


Figure 4.7: Subsystem distribution of the three categorical classes of gene essentiality resulting from the single gene deletion analysis. In total, 193 genes were found to be essential, 230 were partially essential, while the remaining 768 genes were characterized as non-essential.

of exogenous nutrients.

Contrary to the case of the amino acid metabolism, a large amount of metabolic flexibility exists for the reactions associated with lipid metabolism. Only 29 out of 220 genes of the lipid metabolism were found to be essential in the minimal glucose medium. This was primarily due to the existence of complementary isozymes, which was quite prevalent for the associated reactions (around 40%). However, due to severely restricted biochemical information on the encoded enzymes, the specificity of the various enzymes were rather broadly implemented in the model. This will most likely result in an underestimate of genuine gene essentiality, as the overlapping activity of the genes in the model does not accurately reflect the more specialized affinities of the encoded enzymes *in vivo*.

While the ternary gene essentiality analysis provide valid insight into how the specific growth rate is affected by single gene deletions, they do not offer any assessment of the extent of the resulting metabolic flux redistribution. Using the deletion impact p, the quantity of flux redistribution was evaluated for every gene in the model by calculating the emergent flux redistribution following the genetic perturbation. The distribution of p values are presented in Figure 4.8, where the essentiality classifications from the earlier analysis are highlighted.

There is a clear separation of the essential genes showing the largest flux deviations from the wild type network. These genes are clustering at a p value of over $10\text{x}10^7$, cor-

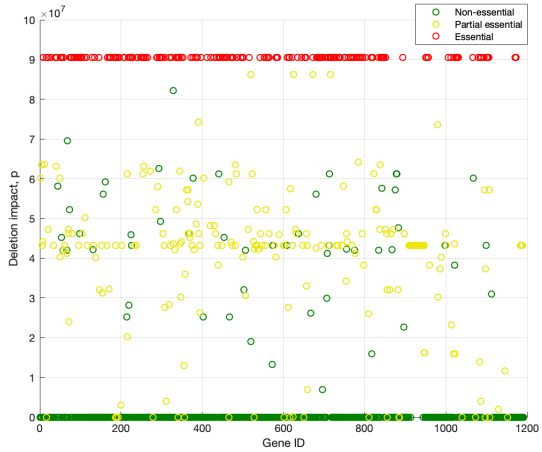
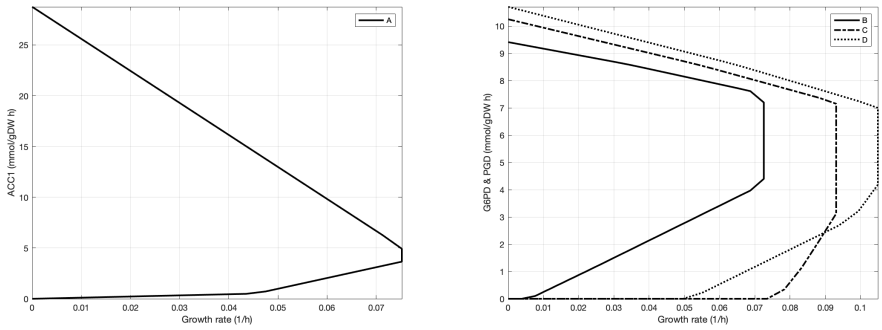


Figure 4.8: The deletion impact p for all 1191 genes of the reconstructed GSM of *A.* sp. T66. The flux distributions were calculated using an FBA formulation of the perturbed network. Also indicated are the ternary essentiality classes resulting from the single gene deletion analysis.

responding to an inactive metabolic network of all zero flux values. Similarly, most of the non-essential genes are clustered at the bottom, where the deletion impact of 0 indicate an identical flux distribution to that of the wild type. However, 50 essential genes show fairly large flux deviations compared to the wild type network. These single gene knockouts indicate that the metabolic network of *A.* sp. T66 in several cases are able to redistribute its metabolic fluxes in response to the genetic perturbation, obtaining the same optimal growth phenotype as the wild type network. This is indicative of an adaptable metabolism, which in a robust fashion redirects its metabolic fluxes to counteract the effect of the genetic knockout, reaching the same optimal growth phenotype as that of the wild type network. The deletion impact of the partially essential genes are rather varied, some showing very large deviations, while others generate flux distributions similar to the wild type.

As in the original paper, as well as that of the aforementioned analysis, the flux distributions of the wild type and mutant networks were all calculated using a standard FBA-formulation of growth optimization. The method does not take alternative optimal solutions into account, whose underlying flux distributions could be quite different from one another. This is especially the case for type III extreme pathways, whose constituent reactions form internally closed loops, able to take on infeasibly large flux values. If these loops are active in a certain flux distribution and not in another, the difference between these may subsequently have a considerable impact on the calculated p value, overestimating the deletion impact of a given gene knockout.

A more appropriate approach could have been to minimize the sum of absolute fluxes [121, 146], then use the resulting flux distributions to calculate the associated deletion impacts. This method is founded on the assumption that there is selection pressure for cells to utilize a minimal amount of enzyme to obtain a given growth phenotype [146]. Conse-



(a) Malonyl-CoA - ACC1 **(b)** NADPH - G6PD and PGD

Figure 4.9: Production envelopes for the *in silico* double reaction mutants proposed by the OptKnock algorithm for increased production of (**a**) malonyl-CoA by ACC1, and increased generation of (**b**) NADPH via the oxidative pathway of the PPP, G6PD and PGD. The graphs indicate the minimal and maximal flux values obtainable for the target reactions at various growth rates.

quently, in addition to providing a unique flux distribution with minimally active type III pathways, the emergent flux distributions may therefore be more biologically realistic, increasing the likelihood of accurately predicting the emergent flux redistribution following a genetic perturbation.

4.3 Model employment for phenotypic predictions

4.3.1 Innate potentiality of the metabolic network uncover strategies for increased production of malonyl-CoA and NADPH

The reconstructed GSM provides a suitable framework for proposing candidate genetic interventions for increased lipid production. Using the OptKnock algorithm, candidate double reaction knockout mutants were found for three of the four considered target reactions (Table 4.7). Depending on the associated gene rules, the suggested mutant strategies required three to four independent gene knockouts. Consequently, in addition to evaluating the calculated yields and production rates, one has to simultaneously consider the tractability of the various genetic interventions.

Malonyl-CoA - ACC1

The first target reaction was ACC1 (AUR0129), which synthesizes the fatty acid precursor malonyl-CoA. The suggested mutant (mutant A) was a double reaction mutant of phosphoglycerate kinase, AUR0008 (PGK, EC:2.7.2.3) and ribose-5-phosphate isomerase, AUR0046 (RPIA, EC:5.3.1.6). Deleting these metabolic reactions from the network resulted in an approximately two-fold increase in the flux through the target reaction, altering its flux variability from 1.89 - 1.89 mmol gDW^{-1} h^{-1} in the wild type network, to 3.66 -

4.93 mmol gDW^{-1} h^{-1} in the knockout mutant (Table 4.7).

This enhanced production of malonyl-CoA was accompanied by a concurrent reduction in the specific growth rate from 0.134 to 0.0753 h^{-1}, indicating that the redirection of carbon flux towards malonyl-CoA production results in an inability to generate appropriate levels of certain biomass precursors. The production envelope (Figure 4.9), which shows the minimal and maximal fluxes of the target reaction for various growth rates, indicate that the rather modest coupling of biomass production and ACC1 flux is not initiated until the growth rate exceeds 0.04 h^{-1}. From this point, the malonyl-CoA production increases gradually until it reaches its maximal capacity at a growth rate of 0.0753 h^{-1}.

This double reaction knockout creates an obstruction in glycolysis, resulting in an increased flux into the PPP. Here, the removal of RPIA induces a redistribution of flux, which eventually leads to an increased production of glycerone phosphate. Through a series of interconversions, the generated glycerone phosphate is transformed into cytosolic pyruvate which subsequently enters the mitochondria, fueling the citric acid cycle, which in the mutant has an increased flux compared to that of the wild type network (citrate

Table 4.7: Suggested double reaction knockout mutants by the OptKnock algorithm for increased productivity of each target metabolite. The list of genes for each mutant strategy indicate the set of genes that need to be disrupted in order to knock out the reaction pairs. Given are the resulting specific growth rates and associated flux ranges through the target reactions. No double reaction knockout strategy were identified for the target reaction ME.

ID	Knockouts	Genes	Growth[a]	Target[b]	Wild type[b]
Malonyl-CoA					
A	AUR0008	T66006855.2	0.0753	3.66 - 4.93	1.89 - 1.89
		T66010974.1			
	AUR0046	T66011653.1			
NADPH, PPP					
B	AUR0008	T66006855.2	0.0727	4.41 - 7.21	0.288 - 0.289
		T66010974.1			
	AUR1205	x[c]			
C	AUR0008	T66006855.2	0.0932	3.13 - 7.17	0.288 - 0.289
		T66010974.1			
	AUR0034	T66006855.2			
		T66010974.1			
D	AUR0008	T66006855.2	0.105	4.19 - 7.02	0.288 - 0.289
		T66010974.1			
	AUR0054	T66004892.1			
NADPH, ME					
-					

[a] Denotes the optimal specific growth rate (h^{-1}).
[b] Predicted flux ranges of the target reactions for the perturbed and wild type network were calculated by FVA at optimum. Given in mmol gDW^{-1} h^{-1}.
[c] Any of the enzymatic subunits of the ubiquinol-cytochrome-c reductase complex: T66000711.1, T66002810.1, T66002825.1, contig_179 or contig_6507.

synthase (CS, EC:2.3.3.1) flux of 1.83 and 2.29 mmol gDW^{-1} h^{-1}, respectively). The generated citrate is then exported out to the cytosol, where it upon the action of ACL is used to generate acetyl-CoA, which subsequently gets carboxylated, forming the end product malonyl-CoA.

Although the algorithm predicts an increased flux through ACC1, the generated malonyl-CoA is predicted to be directly decarboxylated by the action of malonyl-CoA decarboxylase (MCD, EC:4.1.1.9), preventing any increased lipid production. If implemented *in vivo*, it might therefore be necessary to either inhibit the activity of MCD, or knocking out the gene entirely, in order for the proposed mutant to increase its malonyl-CoA productivity.

An even more detrimental side effect of this knockout mutant is the adverse reduction in flux of two of the key NADPH-generating reactions associated with lipid accumulation. The flux variability of the NADPH-producing reactions of the PPP (G6PD and PGD) are in the mutant drastically reduced from that of the wild type network, both showing a more than 10^5 fold reduction. Consequently, the positive effects of the proposed knockout strategy might therefore be quite limited when implemented *in vivo*, questioning its applicability.

NADPH - G6PD and PGD

A joint OptKnock formulation was proposed for the two NADPH-generating reactions of the pentose-phosphate pathway, G6PD (AUR0055) and PGD (AUR0041), as they are subject to a direct flux coupling in *A.* sp. T66 because of their presence in a linear, unbranched pathway. The three top candidate strategies (B, C and D) are presented in Table 4.7, with the resulting production envelopes in Figure 4.9.

The three proposed mutants showed a highly significant increase in flux ranges for the target reactions, compared to that of the wild type network. Common to all the knockout strategies was the inactivation of the glycolytic enzyme PGK (AUR0008), which forces a flux redistribution from the degradational pathway of glycolysis into the oxidative steps of the PPP. This increased influx of carbon causes both of the target reactions to drastically increase their flux capacities, consequently increasing their rate of NADPH production. In fact, by only performing this single-reaction knockout, the minimal and maximal flux values of the target reactions increases to 2.68 and 6.55 mmol gDW^{-1} h^{-1}, respectively, with a concurrent optimal growth rate of 0.106 h^{-1}.

The second knockout was quite different for the three proposed strategies.

The suggested knockouts were reaction AUR1205, ubiquinol-cytochrome-c reductase complex (EC:7.1.1.8) for mutant B, reaction AUR0034, mitochondrial malate dehydrogenase (MDH2, EC:1.1.1.37) for C, and reaction AUR0054, glycerate 2-kinase (GLYCTK, EC:2.7.1.165) for D.

Although the details of the flux redistribution in these mutants are harder to pinpoint, they all induce a drastic increase in flux into the oxidative steps of the PPP, resulting in a considerable loss of carbon because of the oxidative decarboxylation of 6-phospho-d-gluconate by PGD. The generated carbon dioxide is subsequently exported out of the system, preventing its incorporation into the necessary biomass precursors, causing the associated growth rates to decrease. As illustrated in the production envelopes in Figure 4.9, mutant B exhibit the earliest flux coupling between growth and target reactions at a

growth rate of merely 0.005 h^{-1}. On the contrary, mutants C and D do not exhibit any flux coupling until the growth rate has exceeded around 0.05 and 0.07 h^{-1}, respectively.

While all three approaches have a highly positive impact on the metabolic flux through the oxidative steps of the PPP, they do all result in a concurrent decrease in the flux ranges of ACC1. Much like the effect of knockout strategy A on the fluxes through G6PD and PGD, the flux through ACC1 did show a reduction for all three knockouts B, C and D. However, the effect was not as detrimental, only showing a reduction in flux range from 1.88 - 1.89 mmol gDW^{-1} h^{-1} in the unperturbed network, to 0.812 - 0.813 mmol gDW^{-1} h^{-1} in mutant B, 1.04 - 1.04 mmol gDW^{-1} h^{-1} in mutant C, and 1.17 - 1.18 mmol gDW^{-1} h^{-1} in mutant D.

Based on all these results, it seems that knockout strategy D could be the one with the greatest potential, assuming optimal growth. The mutant is showing a small reduction in growth rate, a significant increase in flux through the target reactions, a fairly low reduction in malonyl-CoA production, as well as only needing the simultaneous knockout of three enzyme-encoding genes (Table 4.7). For sub-optimal growth phenotypes, however, the other mutant strategies B and C might be preferable. B would be favorable for growth rates of around 0.07 h^{-1} and below, while the strong growth coupling of mutant C in the growth range 0.08 - 0.09 h^{-1} makes this the best choice at these growth rates.

NADPH - ME

OptKnock did not identity any viable double reaction knockout strategies for increasing the metabolic flux through the cytosolic ME (AUR0135). Additionally, the flux variability of this reaction was only 0 - 0.0002 mmol gDW^{-1} h^{-1} in the wild type network, suggesting that this reaction might not be as important for NADPH generation in *A.* sp. T66 and other oleaginous microorganisms as earlier hypothesized [89].

While the proposed knockout strategies might be accurate when the cells has evolved over time towards optimal growth performance, the flux distribution of the initial generations might be rather different. MOMA was therefore used to calculate the initial flux redistribution following the genetic perturbations, assuming it to be minimally different from that of the wild type network (Table 4.8).

Table 4.8: Flux rates for the target reactions of the four independent double reaction knockout strategies proposed by OptKnock. These fluxes were calculated by MOMA to investigate how the initial flux redistribution of the metabolic network would affect the fluxes through the target reactions.

ID	Target production[a]
A	0.890
B	0.333
C	0.236
D	0.360

[a] Predicted flux of the target reactions, mmol gDW^{-1} h^{-1}.

MOMA predicted that the initial fluxes through the target reactions are quite different from those predicted by OptKnock, and indicated that the beneficial effects of the proposed knockouts would not emerge until the cells have had time to evolve towards a state of optimal growth. Consequently, this leads to the conclusion that the implementation of any of these knockout strategies should be followed by adaptive laboratory evolution of fast-growing strains, ensuring an optimal coupling between the cellular objective and the production of the target metabolite.

A central issue with the OptKnock framework is the implicit assumption of biomass production as an appropriate cellular objective during target metabolite production. In many cases, the biosynthesis of the target compound is not properly initiated during the exponential growth phase, but instead pursues the depletion of an essential nutrient. This nutrient deficiency triggers a metabolic shift with a concurrent redistribution of fluxes within the metabolic network, prompting the generation of the target metabolite. An induced coupling between growth and target metabolite production is therefore quite unrealistic, and would presumably yield poor results when implemented *in vivo*. This uncoupling of growth and target metabolite production is also the case for lipid production in oleaginous microorganisms. Here, the depletion of nitrogen, or some other essential nutrient, induces a substantial redistribution of flux in which lipids are synthesized at high rates during a post-growth lipid accumulation phase [90]. However, experimental efforts have been performed on a *Schizochytrium* strain in which it was found that single-stage continuous cultures at optimal dilution rates could prove beneficial for lipid and DHA production [147]. The reason being factors such as set-up times and harvesting methods, which for the biphasic batch-fermentations were sub-optimal compared to that of the continuous cultures. Consequently, it could therefore be worth exploring these candidate mutant strategies in *A.* sp. T66 to assess whether this could offer novel approaches to increased lipid accumulation in thraustochytrids.

4.3.2 Genome-wide transcriptomic changes are associated with the metabolic shift from growth to lipid accumulation

Integration of the transcriptomics data with the three condition-specific models reveal divergent modes of regulation during the metabolic shift from biomass production to lipid accumulation (Table 4.9). At the onset of lipid accumulation (N1), 188 gene-reaction pairs appears to be regulated on the transcriptional level categorized by a high correlation between the change of their metabolic fluxes and a concurrent change in the associated transcript levels. Out of the 8 major subsystems, the amino acid metabolism appears to be the most significantly regulated on the transcriptional level, constituting 31% of these gene-reaction pairs (Figure 4.10). From the heatmap in Figure 4.11, a considerable proportion of these consist of the biosynthetic aminoacyl-tRNA reactions, suggesting a global reduction in protein synthesis as a response to the ongoing nitrogen depletion. In fact, of the 20 common amino acids, the only tRNA-charging reaction that was not under significant transcriptional regulation was that of cysteinyl-tRNA synthetase (CARS, EC:6.1.1.16). The apparent regulatory mechanisms behind the coinciding reduction in the particular amino acid biosynthetic pathways, however, appears to be more varied.

The biosynthetic pathways of the aromatic amino acids phenylalanine, tyrosine and

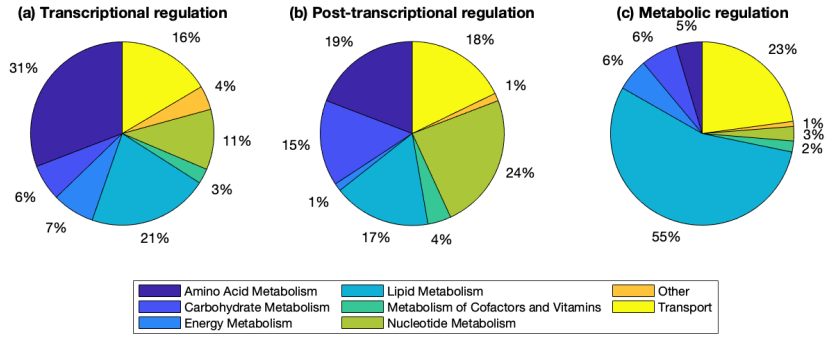


Figure 4.10: Subsystem distribution of the various gene-reaction pairs significantly regulated at the metabolic shift from exponential growth to the early onset of lipid accumulation (N1/E). The modes of regulation are: (a) transcriptional level, showing a high correlation between the differential changes in flux and transcript levels, (b) post-transcriptional level, no change in flux levels with an associated change in transcript levels, (c) metabolic level, inverse correlation between fluxes and transcript levels, or a significant increase in flux with no concurrent change in transcript levels.

tryptophan, in addition to histidine, are highly enriched with transcriptional regulation, showing a considerable downregulation of both metabolic fluxes and associated transcript levels when transitioning into lipid accumulation (Figure 4.11). Interestingly, recent research that applied the same methodology on a GSM of the oleaginous yeast *Y. lipolytica* also suggest that a similar set of metabolic pathways of the amino acid metabolism are subject to transcriptional downregulation during nitrogen limitation [148]. While the identified subsystems consisted of the biosynthetic pathways of histidine, phenylalanine, tyrosine and tryptophan, they also include the biosynthetic pathways of leucine and lysine, which is not found to be significantly regulated transcriptionally in the case of *A.* sp. T66. In the paper, they further hypothesized that the protein TORC1 might possess an alternative regulatory role on lipid accumulation [148], which was further supported by later experimental efforts suggesting an interplay between TORC1, the leucine-intermediate 2-isopropylmalate and a Leu3-like transcription factor [149].

Although the presented results do not fully corroborate these findings, they do point in

Table 4.9: Number of gene-reaction pairs subject to either of the three modes of regulation: transcriptional, post-transcriptional, and metabolic for each of the given condition-comparisons. Lists of the set of genes, associated reactions and associated flux and transcript changes are available in the Supplementary Material 'Transcriptome Model Integration.xlsx'

Condition-comparison	Transcriptional	Post-transcriptional	Metabolic
N1/E	188	148	669
N2/E	190	188	669
N2/N1	103	633	138

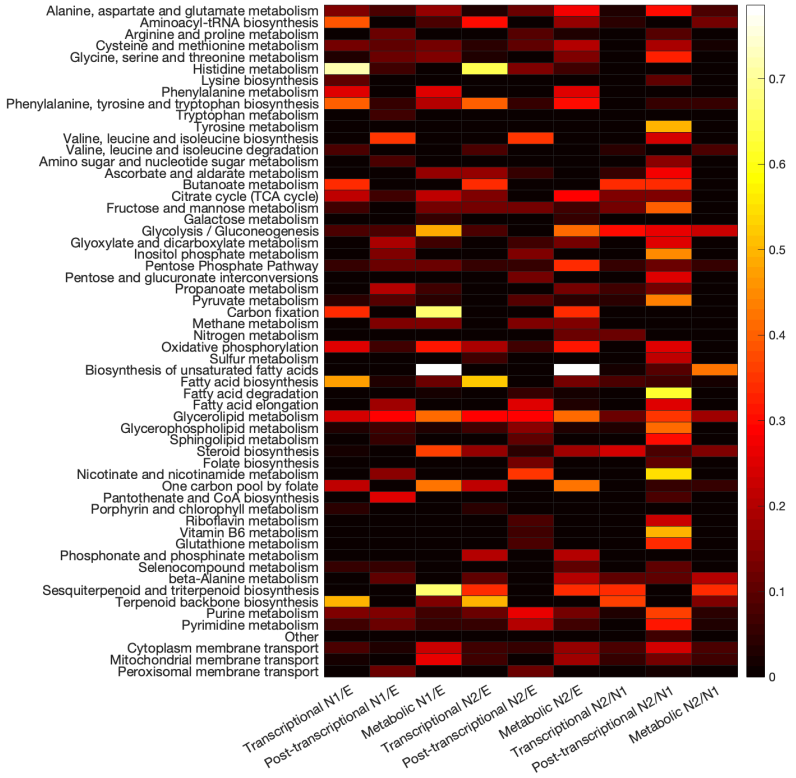


Figure 4.11: Heatmap showing the extent of transcriptional, post-transcriptional and metabolic modes of regulation occurring between the three conditions. The color grading are proportional to the likelihood of a metabolic reaction in a particular subsystem being regulated transcriptionally, post-transcriptionally or metabolically. Abbreviations: E: exponential growth phase, N1: onset of lipid accumulation, N2: late lipid accumulation.

the same direction, alluding to a similar mechanism by which oleaginous microorganisms concertedly downregulates specific subsections of the amino acid metabolism upon nitrogen starvation by way of transcriptional regulation. This could thereby contribute to the rerouting of carbon flow into lipid biosynthesis during the metabolic shift from exponential growth to lipid accumulation.

Whereas lipid metabolism in the aforementioned studies were found to be under limited transcriptional regulation, instead being subject to mostly metabolic regulation [148, 149], our results indicate that the two modes of fatty acid biosynthesis in *A.* sp. T66 are in fact subject to divergent regulatory constraints.

The reactions of the standard fatty acid biosynthetic pathway appear to be under significant transcriptional regulation, showing correlated changes in a large subset of fluxes and associated transcript levels from the exponential phase (E) to the early lipid accumulation

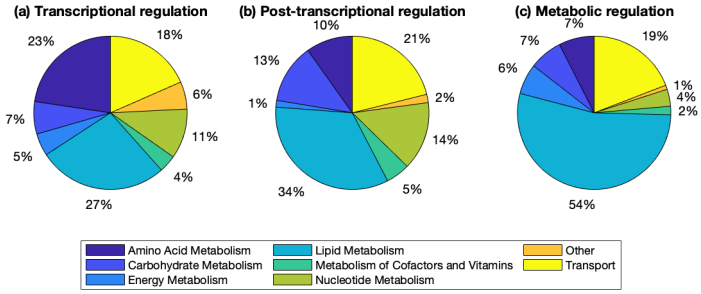


Figure 4.12: Subsystem distribution of the various gene-reaction pairs significantly regulated at the metabolic transition from early to late lipid accumulation (N2/N1). The modes of regulation are: (a) transcriptional level, showing a high correlation between the differential changes in flux and transcript levels, (b) post-transcriptional level, no change in flux levels with an associated change in transcript levels, (c) metabolic level, inverse correlation between fluxes and transcript levels, or a significant increase in flux with no concurrent change in transcript levels.

phase (N1). This regulation is predominately present during the onset of lipid accumulation (N1), while the transcripts level show a minor decrease towards the end of lipid accumulation (N2) (4.11). These reactions are all associated with a single gene encoding both of the catalytic subunits of the FAS enzymatic complex, which shows a significant transcriptional upregulation during the two sample points of the lipid accumulation phase (fold changes of 2.8 and 2.1, respectively). The reactions of the PKS pathway, however, seem to predominately be under metabolic regulation, as seen in Figure 4.11, constituting the majority of the gene-reaction pairs associated with the subsystem 'Biosynthesis of unsaturated fatty acids' (Figure 4.10). The metabolic regulation of the reactions of the PKS pathway seems to be fairly constant during both stages of the lipid accumulation phase.

These contrasting regulatory mechanisms might hint at a possible strategy for increasing both the amount and fractional composition of PUFAs in the lipid moieties of *A.* sp. T66. By increasing the available levels of intracellular lipid precursors (e.g. malonyl-CoA and NADPH), while simultaneously inhibiting the expression of the FAS-encoding gene, one would tentatively proceed to increase the catalytic activity of the PKS synthase pathway, while suppressing the activity of the competing FAS system. In this way, the rate of PUFA biosynthesis would increase, at the same time as the rate of production of the unsaturated fatty acids by the FAS complex would decline. While approaches for the transcriptional inhibition of the FAS-encoding gene might be challenging to develop, other methods such as the direct inhibition of the protein complex using a range of established FAS-inhibitors could also be employed to reduce its cellular activity [150].

The incorporation of the genome-wide transcriptomics data gave rise to valuable insight into the systems-level regulatory mechanisms underlying the metabolic shift into lipid accumulation. However, the results do also hint at a rather simplistic assumption of an exclusive redistribution of metabolic fluxes from biomass production towards lipid accumulation. This is quite apparent when considering the large number of post-transcriptional

gene-reaction pairs associated with the shift from early lipid accumulation (N1) to late lipid accumulation (N2). In total, 633 unique gene-reaction pairs are subject to a significant change in transcript levels, with no concurrent change in metabolic flux (Table 4.9). While a subset of these may correspond to genuine instances of post-transcriptional regulation, the considerable increase in number could rather indicate that the predicted flux distributions might underestimate the global biochemical activity of the metabolic network *in vivo*. In fact, as the subsystem distribution of the associated reactions indicate in Figure 4.12, the underlying transcriptional regulation appears to be generally representative of the subsystem distribution of the metabolic network of *A.* sp. T66 as a whole (Figure 4.5). This suggests a comprehensive transcriptional transition not reflected in the metabolic fluxes of the model, which could be indicative of a metabolic network whose global activity are underestimated due to the condition-dependent assumptions.

Chapter 5

Conclusion and Outlook

A high-quality GSM of the entire metabolism of the thraustochytrid *A.* sp. T66 was successfully reconstructed. The model termed iVS1191 was subject to extensive manual curation and refinements, considerably expanding its metabolic scope to that of the template reconstruction iCY1170_DHA. In doing so, the connectivity of the metabolic network showed drastic improvements when compared to iCY1170_DHA, significantly reducing the number of blocked reactions and associated dead-end metabolites.

Through model employment using the OptKnock framework, multiple double reaction knockout strategies were identified which were predicted to increase the productivity of the essential PUFA precursors malonyl-CoA and NADPH. Furthermore, integration of genome-wide transcriptomics data from fermentation experiments revealed a transcriptional downregulation of subsets of the metabolic fluxes associated with the amino acid metabolism, alluding to a conserved regulatory mechanism also observed in other oleaginous microorganisms. This analysis also suggested a bimodal regulatory scheme in which the FAS complex appears to predominately be regulated on the transcriptional level, while the competing PKS pathway seems to be under metabolic control. This led to the model-generated hypothesis that increased levels of lipid precursors might preferentially influence the activity of the PKS pathway, while in minor ways affecting that of the FAS system, thus increasing the fractional composition of PUFAs.

This thesis successfully delivered on both the primary and secondary aim of model reconstruction and application for identifying strategies for increased production of PUFA-containing lipids. However, the general lack of validations of the model predictions is of a genuine concern. Before future employments of the reconstruction, it would be highly beneficial to perform large-scale validations of the phenotypic predictions. These could be in the form of the growth predictions performed during this project or by utilizing high-throughput phenotype microarrays on a wide range of nutrient sources. The resulting disparities between predicted and experimental results could thereafter be used to drive future model refinements.

Fundamental to the validation of these predictions is detailed quantitative data on the biomass composition and associated energy demands of *A.* sp. T66. As the analysis in this

project indicates, the current biomass composition and associated energy demands are of insufficient quality. Obtaining high-quality data should therefore be of the highest priority before future validations and applications of the reconstruction. Planning of these experiments are currently being discussed by the collaborators of this project (AurOmega).

Although interesting in their own right, the gene essentiality predictions should primarily be utilized for model validation when *in vivo* knockout data is available. Disparities between the experimental and predicted mutant phenotypes will subsequently provide indications of inadequacies in the reconstruction, prompting future model refinements.

The model-generated hypothesis of the bimodal regulation of FAS and PKS is a great starting point for further research. A reasonable approach is to investigate whether the changing proteomic levels of these complexes are correlated with the corresponding transcript changes. If this turns out to be the case, the subsequent plan of attack could be the identification of strategies for increasing the lipid precursor pool. Ways of increasing these may be difficult to ascertain, although crudely over-expressing enzymes such as ACC1 could prove beneficial. A more interesting and possibly more promising strategy might be to exploit the innate biomolecular modes of regulation responsible for the metabolic shift from exponential growth to lipid accumulation. As our results indicate, the downregulation of specific subsets of the amino acid metabolism appear to be closely linked to the rerouting of flux towards lipid production. Further investigation in the form of differential gene co-expression analysis could aid in the identification of key regulators that are responsible for this transcriptional and metabolic shift, which could highlight novel strategies for increasing the capabilities of lipid biosynthesis in *A.* sp. T66.

Bibliography

[1] Ryota Hosomi, Munehiro Yoshida, and Kenji Fukunaga. Seafood consumption and components for health. *Global journal of health science*, 4(3):72–86, 2012.

[2] Danielle Swanson, Robert Block, and Shaker A. Mousa. Omega-3 fatty acids epa and dha: health benefits throughout life. *Advances in nutrition (Bethesda, Md.)*, 3(1):1–7, 2012.

[3] Lotte Lauritzen, Paolo Brambilla, Alessandra Mazzocchi, Laurine B. S. Harsløf, Valentina Ciappolino, and Carlo Agostoni. Dha effects in brain development and function. *Nutrients*, 8(1):6, 2016.

[4] Jaclyn M. Coletta, Stacey J. Bell, and Ashley S. Roman. Omega-3 fatty acids and pregnancy. *Reviews in obstetrics gynecology*, 3(4):163–171, 2010.

[5] Richard J. Bloomer, Douglas E. Larson, Kelsey H. Fisher-Wellman, Andrew J. Galpin, and Brian K. Schilling. Effect of eicosapentaenoic and docosahexaenoic acid on resting and exercise-induced inflammatory and oxidative stress biomarkers: a randomized, placebo controlled, cross-over study. *Lipids in health and disease*, 8:36–36, 2009.

[6] M. Bouwens, O. van de Rest, N. Dellschaft, M. G. Bromhaar, L. C. de Groot, J. M. Geleijnse, M. Muller, and L. A. Afman. Fish-oil supplementation induces antiinflammatory gene expression profiles in human blood mononuclear cells. *Am J Clin Nutr*, 90(2):415–24, 2009.

[7] The state of world fisheries and aquaculture 2018 - meeting the sustainable development goals. Report, Food and Agriculture Organization of the United Nations, 2018.

[8] Ian H. Pike and Andrew Jackson. *Fish oil: Production and use now and in the future*, volume 22. 2010.

[9] D. F. Rawn, D. S. Forsyth, J. J. Ryan, K. Breakell, V. Verigin, H. Nicolidakis, S. Hayward, P. Laffey, and H. B. Conacher. Pcb, pcdd and pcdf residues in fin

and non-fin fish products from the canadian retail market 2002. *Sci Total Environ*, 359(1-3):101–10, 2006.

[10] Kh M. El-Moselhy, A. I. Othman, H. Abd El-Azem, and M. E. A. El-Metwally. Bioaccumulation of heavy metals in some tissues of fish in the red sea, egypt. *Egyptian Journal of Basic and Applied Sciences*, 1(2):97–105, 2014.

[11] A. Gupta, C. J. Barrow, and M. Puri. Omega-3 biotechnology: Thraustochytrids as a novel source of omega-3 oils. *Biotechnol Adv*, 30(6):1733–45, 2012.

[12] B. Leyland, S. Leu, and S. Boussiba. Are thraustochytrids algae? *Fungal Biol*, 121(10):835–840, 2017.

[13] J. Li, R. Liu, G. Chang, X. Li, M. Chang, Y. Liu, Q. Jin, and X. Wang. A strategy for the highly efficient production of docosahexaenoic acid by aurantiochytrium limacinum sr21 using glucose and glycerol as the mixed carbon sources. *Bioresour Technol*, 177:51–7, 2015.

[14] T. Y. Huang, W. C. Lu, and I. M. Chu. A fermentation strategy for producing docosahexaenoic acid in aurantiochytrium limacinum sr21 and increasing c22:6 proportions in total fatty acid. *Bioresour Technol*, 123:8–14, 2012.

[15] Joris Beld, D. John Lee, and Michael D. Burkart. Fatty acid biosynthesis revisited: structure elucidation and metabolic engineering. *Molecular bioSystems*, 11(1):38–59, 2015.

[16] C. Ratledge. Fatty acid biosynthesis in microorganisms being used for single cell oil production. *Biochimie*, 86(11):807–15, 2004.

[17] Edward J. O'Brien, Jonathan M. Monk, and Bernhard O. Palsson. Using genome-scale models to predict biological capabilities. *Cell*, 161(5):971–987, 2015.

[18] I. Thiele and BØ Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5(1):93–121, 2010.

[19] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

[20] C. Ye, W. Qiao, X. Yu, X. Ji, H. Huang, J. L. Collier, and L. Liu. Reconstruction and analysis of the genome-scale metabolic model of schizochytrium limacinum sr21 for docosahexaenoic acid production. *BMC Genomics*, 16:799, 2015.

[21] Han-Yu Chuang, Matan Hofree, and Trey Ideker. A decade of systems biology. *Annual review of cell and developmental biology*, 26:721–744, 2010.

[22] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–72, 2001.

[23] D. Noble. The rise of computational biology. *Nat Rev Mol Cell Biol*, 3(6):459–63, 2002.

[24] Frederick S. Hillier and Gerald J. Lieberman. *Introduction to operations research*. 2015.

[25] J. Lundgren, M. Rönnqvist, and P. Värbrand. *Optimization*. Professional Publishing, 2010.

[26] Wikimedia Commons. Ein dreidimensionales, konvexes und beschränktes polyeder, 2008.

[27] George B. Dantzig. *Origins of the simplex method*, pages 141–151. ACM, 1990.

[28] Victor Klee and George Minty. *How good is the simplex algorithm?*, volume III, pages 159–175. Academic Press, 1972.

[29] A. Forsgren, P. Gill, and M. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 44(4):525–597, 2002.

[30] Jacek Gondzio, Tam, 225, and s Terlaky. *A computational view of interior point methods*, pages 103–144. Oxford University Press, Inc., 1996.

[31] E. O. Voit. The best models of metabolism. *Wiley Interdiscip Rev Syst Biol Med*, 9(6), 2017.

[32] W. W. Chen, M. Niepel, and P. K. Sorger. Classic and contemporary approaches to modeling biochemical reactions. *Genes Dev*, 24(17):1861–75, 2010.

[33] M. A. Oberhardt, B. O. Palsson, and J. A. Papin. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*, 5:320, 2009.

[34] A. Trewavas. A brief history of systems biology. "every object that biology studies is a system of systems." francois jacob (1974). *Plant Cell*, 18(10):2420–30, 2006.

[35] O. D. Kim, M. Rocha, and P. Maia. A review of dynamic modeling approaches and their application in computational strain optimization for metabolic engineering. *Front Microbiol*, 9:1690, 2018.

[36] E. Vasilakou, D. Machado, A. Theorell, I. Rocha, K. Noh, M. Oldiges, and S. A. Wahl. Current state and challenges for dynamic metabolic modeling. *Curr Opin Microbiol*, 33:97–104, 2016.

[37] Sergio Grimbs, Joachim Selbig, Sascha Bulik, Hermann-Georg Holzhütter, and Ralf Steuer. The stability and robustness of metabolic states: identifying stabilizing sites in metabolic networks. *Molecular systems biology*, 3:146–146, 2007.

[38] Timo R. Maarleveld, Ruchir A. Khandelwal, Brett G. Olivier, Bas Teusink, and Frank J. Bruggeman. Basic concepts and principles of stoichiometric modeling of metabolic networks. *Biotechnology journal*, 8(9):997–1008, 2013.

[39] L. Kuepfer. Stoichiometric modelling of microbial metabolism. *Methods Mol Biol*, 1191:3–18, 2014.

[40] G. Stephanopoulos. Metabolic fluxes and metabolic engineering. *Metab Eng*, 1(1):1–11, 1999.

[41] N. E. Lewis, H. Nagarajan, and B. O. Palsson. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol*, 10(4):291–305, 2012.

[42] Bernhard Ø Palsson. *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge University Press, Cambridge, 2015.

[43] R. Schuetz, L. Kuepfer, and U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli. *Mol Syst Biol*, 3:119, 2007.

[44] J. S. Edwards, R. U. Ibarra, and B. O. Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol*, 19(2):125–30, 2001.

[45] R. Mahadevan and C. H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng*, 5(4):264–76, 2003.

[46] D. Segre, D. Vitkup, and G. M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A*, 99(23):15112–7, 2002.

[47] G. J. Baart and D. E. Martens. Genome-scale metabolic models: reconstruction and analysis. *Methods Mol Biol*, 799:107–26, 2012.

[48] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–5, 2002.

[49] S. R. Paladugu, V. Chickarmane, A. Deckard, J. P. Frumkin, M. McCormack, and H. M. Sauro. In silico evolution of functional modules in biochemical networks. *Syst Biol (Stevenage)*, 153(4):223–35, 2006.

[50] T. Y. Kim, S. B. Sohn, Y. B. Kim, W. J. Kim, and S. Y. Lee. Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr Opin Biotechnol*, 23(4):617–23, 2012.

[51] E. Simeonidis and N. D. Price. Genome-scale modeling for metabolic engineering. *J Ind Microbiol Biotechnol*, 42(3):327–38, 2015.

[52] J. P. Faria, M. Rocha, I. Rocha, and C. S. Henry. Methods for automated genome-scale metabolic model reconstruction. *Biochem Soc Trans*, 46(4):931–936, 2018.

[53] Rasmus Agren, Liming Liu, Saeed Shoaie, Wanwipa Vongsangnak, Intawat Nookaew, and Jens Nielsen. The raven toolbox and its use for generating a genome-scale metabolic model for penicillium chrysogenum. *PLOS Computational Biology*, 9(3):e1002980, 2013.

[54] H. Wang, S. Marcisauskas, B. J. Sanchez, I. Domenzain, D. Hermansson, R. Agren, J. Nielsen, and E. J. Kerkhoven. Raven 2.0: A versatile toolbox for metabolic network reconstruction and a case study on streptomyces coelicolor. *PLoS Comput Biol*, 14(10):e1006541, 2018.

[55] Christopher S. Henry, Matthew DeJongh, Aaron A. Best, Paul M. Frybarger, Ben Linsay, and Rick L. Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28:977, 2010.

[56] William R. Pearson. An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, Chapter 3:Unit3.1–Unit3.1, 2013.

[57] Markus A. Keller, Gabriel Piedrafita, and Markus Ralser. The widespread role of non-enzymatic reactions in cellular metabolism. *Current Opinion in Biotechnology*, 34:153–161, 2015.

[58] A. R. Joyce and B. O. Palsson. Predicting gene essentiality using genome-scale in silico models. *Methods Mol Biol*, 416:433–57, 2008.

[59] Anamika Kothari, Aya Kubo, Carol A. Fulcher, Daniel S. Weaver, Deepika Weerasinghe, Hartmut Foerster, Ingrid M. Keseler, Kate Dreher, Lukas A. Mueller, Mario Latendresse, Markus Krummenacker, Pallavi Subhraveti, Peifen Zhang, Quang Ong, Richard Billington, Ron Caspi, Suzanne Paley, Timothy A. Holland, Tomer Altman, and Peter D. Karp. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 42(D1):D459–D471, 2013.

[60] William Martin. Evolutionary origins of metabolic compartmentalization in eukaryotes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1541):847–855, 2010.

[61] A. Zecchin, P. C. Stapor, J. Goveia, and P. Carmeliet. Metabolic pathway compartmentalization: an underappreciated opportunity? *Curr Opin Biotechnol*, 34:73–81, 2015.

[62] H. Owji, N. Nezafat, M. Negahdaripour, A. Hajiebrahimi, and Y. Ghasemi. A comprehensive review of signal peptides: Structure, roles, and applications. *Eur J Cell Biol*, 97(6):422–441, 2018.

[63] Pierre Dönnes and Annette Höglund. Predicting protein subcellular localization: past, present, and future. *Genomics, proteomics bioinformatics*, 2(4):209–215, 2004.

[64] S. Karniely and O. Pines. Single translation–dual destination: mechanisms of dual protein targeting in eukaryotes. *EMBO Rep*, 6(5):420–5, 2005.

[65] N. Linka and A. P. Weber. Intracellular metabolite transporters in plants. *Mol Plant*, 3(1):21–53, 2010.

[66] Nitish K. Mishra, Junil Chang, and Patrick X. Zhao. Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PLOS ONE*, 9(6):e100278, 2014.

[67] Adam M. Feist and Bernhard O. Palsson. The biomass objective function. *Current opinion in microbiology*, 13(3):344–349, 2010.

[68] Siu H. J. Chan, Jingyi Cai, Lin Wang, Margaret N. Simons-Senftle, and Costas D. Maranas. Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. *Bioinformatics*, 33(22):3603–3609, 2017.

[69] E. Ashley Beck, A. Kristopher Hunt, and P. Ross Carlson. Measuring cellular biomass composition for computational biology applications. *Processes*, 6(5), 2018.

[70] L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdottir, J. Wachowiak, S. M. Keating, V. Vlasov, S. Magnusdottir, C. Y. Ng, G. Preciat, A. Zagare, S. H. J. Chan, M. K. Aurich, C. M. Clancy, J. Modamio, J. T. Sauls, A. Noronha, A. Bordbar, B. Cousins, D. C. El Assal, L. V. Valcarcel, I. Apaolaza, S. Ghaderi, M. Ahookhosh, M. Ben Guebila, A. Kostromins, N. Sompairac, H. M. Le, D. Ma, Y. Sun, L. Wang, J. T. Yurkovich, M. A. P. Oliveira, P. T. Vuong, L. P. El Assal, I. Kuperstein, A. Zinovyev, H. S. Hinton, W. A. Bryant, F. J. Aragon Artacho, F. J. Planes, E. Stalidzans, A. Maass, S. Vempala, M. Hucka, M. A. Saunders, C. D. Maranas, N. E. Lewis, T. Sauter, B. O. Palsson, I. Thiele, and R. M. T. Fleming. Creation and analysis of biochemical constraint-based models using the cobra toolbox v.3.0. *Nat Protoc*, 14(3):639–702, 2019.

[71] S. Klamt, J. Saez-Rodriguez, and E. D. Gilles. Structural and functional analysis of cellular networks with cellnetanalyzer. *BMC Syst Biol*, 1:2, 2007.

[72] S. Pan and J. L. Reed. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. *Curr Opin Biotechnol*, 51:103–108, 2018.

[73] J. D. Orth and B. O. Palsson. Systematizing the generation of missing metabolic knowledge. *Biotechnol Bioeng*, 107(3):403–12, 2010.

[74] I. Thiele, N. Vlassis, and R. M. Fleming. fastgapfill: efficient gap filling in metabolic networks. *Bioinformatics*, 30(17):2529–31, 2014.

[75] M. N. Benedict, M. B. Mundy, C. S. Henry, N. Chia, and N. D. Price. Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models. *PLoS Comput Biol*, 10(10):e1003882, 2014.

[76] Z. Hosseini and S. A. Marashi. Discovering missing reactions of metabolic networks by using gene co-expression data. *Sci Rep*, 7:41774, 2017.

[77] P. D. Karp, D. Weaver, and M. Latendresse. How accurate is automated gap filling of metabolic models? *BMC Syst Biol*, 12(1):73, 2018.

[78] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[79] Claus Jonathan Fritzemeier, Daniel Hartleb, Balázs Szappanos, Balázs Papp, and Martin J. Lercher. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLOS Computational Biology*, 13(4):e1005494, 2017.

[80] Nathan D. Price, Ines Thiele, and Bernhard Ø Palsson. Candidate states of helicobacter pylori's genome-scale metabolic network upon application of "loop law" thermodynamic constraints. *Biophysical journal*, 90(11):3919–3928, 2006.

[81] S. Y. Lee, D. Y. Lee, and T. Y. Kim. Systems biotechnology for strain improvement. *Trends Biotechnol*, 23(7):349–58, 2005.

[82] Lauren B. Pickens, Yi Tang, and Yit-Heng Chooi. Metabolic engineering for the production of natural products. *Annual review of chemical and biomolecular engineering*, 2:211–236, 2011.

[83] R. Agren, J. M. Otero, and J. Nielsen. Genome-scale modeling enables metabolic engineering of saccharomyces cerevisiae for succinic acid production. *J Ind Microbiol Biotechnol*, 40(7):735–47, 2013.

[84] John Blazeck and Hal Alper. Systems metabolic engineering: genome-scale models and beyond. *Biotechnology journal*, 5(7):647–659, 2010.

[85] S. Y. Lee, S. H. Hong, and S. Y. Moon. In silico metabolic pathway analysis and design: succinic acid production by metabolically engineered escherichia coli as an example. *Genome Inform*, 13:214–23, 2002.

[86] H. Alper, Y. S. Jin, J. F. Moxley, and G. Stephanopoulos. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in escherichia coli. *Metab Eng*, 7(3):155–64, 2005.

[87] C. Bro, B. Regenberg, J. Forster, and J. Nielsen. In silico aided metabolic engineering of saccharomyces cerevisiae for improved bioethanol production. *Metab Eng*, 8(2):102–11, 2006.

[88] J. G. Metz, P. Roessler, D. Facciotti, C. Levering, F. Dittrich, M. Lassner, R. Valentine, K. Lardizabal, F. Domergue, A. Yamada, K. Yazawa, V. Knauf, and J. Browse. Production of polyunsaturated fatty acids by polyketide synthases in both prokaryotes and eukaryotes. *Science*, 293(5528):290–3, 2001.

[89] J. P. Wynn, A. bin Abdul Hamid, and C. Ratledge. The role of malic enzyme in the regulation of lipid accumulation in filamentous fungi. *Microbiology*, 145 (Pt 8):1911–7, 1999.

[90] A. N. Jakobsen, I. M. Aasen, K. D. Josefsen, and A. R. Strom. Accumulation of docosahexaenoic acid-rich lipid in thraustochytrid aurantiochytrium sp. strain t66: effects of n and p starvation and o2 limitation. *Appl Microbiol Biotechnol*, 80(2):297–306, 2008.

[91] Z. Li, X. Chen, J. Li, T. Meng, L. Wang, Z. Chen, Y. Shi, X. Ling, W. Luo, D. Liang, Y. Lu, Q. Li, and N. He. Functions of pks genes in lipid synthesis of schizochytrium sp. by gene disruption and metabolomics analysis. *Mar Biotechnol (NY)*, 20(6):792–802, 2018.

[92] Matlab optimization toolbox, 2017b.

[93] Python 3.6.4, 2019.

[94] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke. Cobrapy: Constraints-based reconstruction and analysis for python. *BMC Syst Biol*, 7:74, 2013.

[95] Nikolay Martyushenko and Eivind Almaas. Modelexplorer - software for visual inspection and inconsistency correction of genome-scale metabolic reconstructions. *BMC bioinformatics*, 20(1):56–56, 2019.

[96] Z. A. King, A. Drager, A. Ebrahim, N. Sonnenschein, N. E. Lewis, and B. O. Palsson. Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS Comput Biol*, 11(8):e1004321, 2015.

[97] Gurobi 7.5.2, 2019.

[98] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

[99] L. B. Koski and G. B. Golding. The closest blast hit is often not the nearest neighbor. *J Mol Evol*, 52(6):540–2, 2001.

[100] A. Pertsemlidis and 3rd Fondon, J. W. Having a blast with bioinformatics (and avoiding blastphemy). *Genome Biol*, 2(10):Reviews2002, 2001.

[101] C. A. Kerfeld and K. M. Scott. Using blast to teach "e-value-tionary" concepts. *PLoS Biol*, 9(2):e1001014, 2011.

[102] S. R. Eddy. Where did the blosum62 alignment score matrix come from? *Nat Biotechnol*, 22(8):1035–6, 2004.

[103] Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L. Madden. Ncbi blast: a better web interface. *Nucleic acids research*, 36(Web Server issue):W5–W9, 2008.

[104] B. Gschloessl, Y. Guermeur, and J. M. Cock. Hectar: a method to predict subcellular targeting in heterokonts. *BMC Bioinformatics*, 9:393, 2008.

[105] O. Kilian and P. G. Kroth. Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. *Plant J*, 41(2):175–83, 2005.

[106] J. J. Almagro Armenteros, C. K. Sonderby, S. K. Sonderby, H. Nielsen, and O. Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

[107] Jgi aurli1, 2017.

[108] Pre-trained hmm, eukaryota, identity 100 %, 2017.

[109] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 32(Database issue):D115–9, 2004.

[110] V. C. Goswitz and R. J. Brooker. Structural features of the uniporter/symporter/antiporter superfamily. *Protein science : a publication of the Protein Society*, 4(3):534–537, 1995.

[111] F. Palmieri and C. L. Pierri. Mitochondrial metabolite transport. *Essays Biochem*, 47:37–52, 2010.

[112] Jr. Saier, M. H., V. S. Reddy, B. V. Tsu, M. S. Ahmed, C. Li, and G. Moreno-Hagelsieb. The transporter classification database (tcdb): recent advances. *Nucleic Acids Res*, 44(D1):D372–9, 2016.

[113] T. Gabaldon. Peroxisome diversity and evolution. *Philos Trans R Soc Lond B Biol Sci*, 365(1541):765–73, 2010.

[114] S. G. Gould, G. A. Keller, and S. Subramani. Identification of a peroxisomal targeting signal at the carboxy terminus of firefly luciferase. *J Cell Biol*, 105(6 Pt 2):2923–31, 1987.

[115] C. Brocard and A. Hartig. Peroxisome targeting signal 1: is it really a simple tripeptide? *Biochim Biophys Acta*, 1763(12):1565–73, 2006.

[116] Loris Fossier Marchan, Kim J. Lee Chang, Peter D. Nichols, Jane L. Polglase, Wilfrid J. Mitchell, and Tony Gutierrez. Screening of new british thraustochytrids isolates for docosahexaenoic acid (dha) production. *Journal of applied phycology*, 29(6):2831–2843, 2017.

[117] C. J. Fritzemeier, D. Hartleb, B. Szappanos, B. Papp, and M. J. Lercher. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLoS Comput Biol*, 13(4):e1005494, 2017.

[118] N. D. Price, I. Famili, D. A. Beard, and B. O. Palsson. Extreme pathways and kirchhoff's second law. *Biophys J*, 83(5):2879–82, 2002.

[119] R. B. Stein and J. J. Blum. On the analysis of futile cycles in metabolism. *J Theor Biol*, 72(3):487–522, 1978.

[120] H. Qian and D. A. Beard. Metabolic futile cycles and their functions: a systems analysis of energy and control. *Syst Biol (Stevenage)*, 153(4):192–200, 2006.

[121] A. A. Desouki, F. Jarre, G. Gelius-Dietrich, and M. J. Lercher. Cyclefreeflux: efficient removal of thermodynamically infeasible loops from flux distributions. *Bioinformatics*, 31(13):2159–65, 2015.

[122] I. Thiele, T. D. Vo, N. D. Price, and B. O. Palsson. Expanded metabolic reconstruction of helicobacter pylori (iit341 gsm/gpr): an in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol*, 187(16):5818–30, 2005.

[123] Zixiang Xu, Xiao Sun, and Shihai Yu. Genome-scale analysis to the impact of gene deletion on the metabolism of e. coli: constraint-based simulation approach. *BMC bioinformatics*, 10 Suppl 1(Suppl 1):S62–S62, 2009.

[124] A. P. Burgard, P. Pharkya, and C. D. Maranas. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng*, 84(6):647–57, 2003.

[125] Sergio Bordel, Rasmus Agren, and Jens Nielsen. Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLOS Computational Biology*, 6(7):e1000859, 2010.

[126] I. M. Aasen, T. M. B. Heggeset, H. Ertesvåg, and B. Liu. Lipid accumulation and dha-production in aurantiochytrium sp – responses to nitrogen starvation revealed by analyses of growth and production kinetics and global transcriptomes. Unpublished, N.D.

[127] W. De-Eknamkul and B. Potduang. Biosynthesis of beta-sitosterol and stigmasterol in croton sublyratus proceeds via a mixed origin of isoprene units. *Phytochemistry*, 62(3):389–98, 2003.

[128] J. C. Xavier, K. R. Patil, and I. Rocha. Metabolic models and gene essentiality data reveal essential and conserved metabolism in prokaryotes. *PLoS Comput Biol*, 14(11):e1006556, 2018.

[129] D. Bui, D. Ravasz, and C. Chinopoulos. The effect of 2-ketobutyrate on mitochondrial substrate-level phosphorylation. *Neurochem Res*, 2019.

[130] R. Bolli, K. A. Nalecz, and A. Azzi. Monocarboxylate and alpha-ketoglutarate carriers from bovine heart mitochondria. purification by affinity chromatography on immobilized 2-cyano-4-hydroxycinnamate. *J Biol Chem*, 264(30):18024–30, 1989.

[131] G. B. Kohlhaw. Leucine biosynthesis in fungi: entering metabolism through the back door. *Microbiol Mol Biol Rev*, 67(1):1–15, table of contents, 2003.

[132] N. J. Watmough and F. E. Frerman. The electron transfer flavoprotein: ubiquinone oxidoreductases. *Biochim Biophys Acta*, 1797(12):1910–6, 2010.

[133] D. Chipman, Z. Barak, and J. V. Schloss. Biosynthesis of 2-aceto-2-hydroxy acids: acetolactate synthases and acetohydroxyacid synthases. *Biochim Biophys Acta*, 1385(2):401–19, 1998.

[134] J. Levering, J. Broddrick, C. L. Dupont, G. Peers, K. Beeri, J. Mayers, A. A. Gallina, A. E. Allen, B. O. Palsson, and K. Zengler. Genome-scale model reveals metabolic basis of biomass partitioning in a model diatom. *PLoS One*, 11(5):e0155038, 2016.

[135] D. Meesapyodsuk and X. Qiu. Biosynthetic mechanism of very long chain polyunsaturated fatty acids in thraustochytrium sp. 26185. *J Lipid Res*, 57(10):1854–1864, 2016.

[136] P. Stover and V. Schirch. Serine hydroxymethyltransferase catalyzes the hydrolysis of 5,10-methenyltetrahydrofolate to 5-formyltetrahydrofolate. *J Biol Chem*, 265(24):14227–33, 1990.

[137] P. Stover and V. Schirch. Enzymatic mechanism for the hydrolysis of 5,10-methenyltetrahydropteroylglutamate to 5-formyltetrahydropteroylglutamate by serine hydroxymethyltransferase. *Biochemistry*, 31(7):2155–64, 1992.

[138] E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, A. Drager, N. Mih, F. Gatto, A. Nilsson, G. A. Preciat Gonzalez, M. K. Aurich, A. Prlic, A. Sastry, A. D. Danielsdottir, A. Heinken, A. Noronha, P. W. Rose, S. K. Burley, R. M. T. Fleming, J. Nielsen, I. Thiele, and B. O. Palsson. Recon3d enables a three-dimensional view of gene variation in human metabolism. *Nat Biotechnol*, 36(3):272–281, 2018.

[139] E. R. Kunji. The role and structure of mitochondrial carriers. *FEBS Lett*, 564(3):239–44, 2004.

[140] E. R. Kunji and A. J. Robinson. Coupling of proton and substrate translocation in the transport cycle of mitochondrial carriers. *Curr Opin Struct Biol*, 20(4):440–7, 2010.

[141] Bin Liu, Helga Ertesvåg, Inga Marie Aasen, Olav Vadstein, Trygve Brautaset, and Tonje Marita Bjerkan Heggeset. Draft genome sequence of the docosahexaenoic acid producing thraustochytrid aurantiochytrium sp. t66. *Genomics data*, 8:115–116, 2016.

[142] Chao Ye, Nan Xu, Haiqin Chen, Yong Q. Chen, Wei Chen, and Liming Liu. Reconstruction and analysis of a genome-scale metabolic model of the oleaginous fungus mortierella alpina. *BMC systems biology*, 9:1–1, 2015.

[143] J. Kim, M. Fabris, G. Baart, M. K. Kim, A. Goossens, W. Vyverman, P. G. Falkowski, and D. S. Lun. Flux balance analysis of primary metabolism in the diatom phaeodactylum tricornutum. *Plant J*, 85(1):161–76, 2016.

[144] Joana C. Xavier, Kiran Raosaheb Patil, and Isabel Rocha. Metabolic models and gene essentiality data reveal essential and conserved metabolism in prokaryotes. *PLoS computational biology*, 14(11):e1006556–e1006556, 2018.

[145] Seshagiri Raghukumar. Ecology of the marine protists, the labyrinthulomycetes (thraustochytrids and labyrinthulids). *European Journal of Protistology*, 38(2):127–145, 2002.

[146] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. Konig, R. D. Smith, and B. O. Palsson. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol*, 6:390, 2010.

[147] E. Ganuza and M. S. Izquierdo. Lipid accumulation in schizochytrium g13/2s produced in continuous culture. *Appl Microbiol Biotechnol*, 76(5):985–90, 2007.

[148] Eduard J. Kerkhoven, Kyle R. Pomraning, Scott E. Baker, and Jens Nielsen. Regulation of amino-acid metabolism controls flux to lipid accumulation in yarrowia lipolytica. *Npj Systems Biology And Applications*, 2:16005, 2016.

[149] E. J. Kerkhoven, Y. M. Kim, S. Wei, C. D. Nicora, T. L. Fillmore, S. O. Purvine, B. J. Webb-Robertson, R. D. Smith, S. E. Baker, T. O. Metz, and J. Nielsen. Leucine biosynthesis is involved in regulating high lipid accumulation in yarrowia lipolytica. *MBio*, 8(3), 2017.

[150] J. Wang, R. Hudson, and H. O. Sintim. Inhibitors of fatty acid synthesis in prokaryotes and eukaryotes as anti-infective, anticancer and anti-obesity drugs. *Future Med Chem*, 4(9):1113–51, 2012.

[151] Nadine Pollak, Christian Dölle, and Mathias Ziegler. The power to reduce: pyridine nucleotides–small molecules with a multitude of functions. *The Biochemical journal*, 402(2):205–218, 2007.

[152] Alton Meister. *[1] Glutathione metabolism*, volume 251, pages 3–7. Academic Press, 1995.

[153] Vivi Joosten and Willem J. H. van Berkel. Flavoenzymes. *Current Opinion in Chemical Biology*, 11(2):195–202, 2007.

[154] Wolf-Dieter Lienhart, Venugopal Gudipati, and Peter Macheroux. The human flavoproteome. *Archives of biochemistry and biophysics*, 535(2):150–162, 2013.

[155] J. J. Kim and R. Miura. Acyl-coa dehydrogenases and acyl-coa oxidases. structural basis for mechanistic similarities and differences. *Eur J Biochem*, 271(3):483–93, 2004.

[156] L. Galdieri, T. Zhang, D. Rogerson, R. Lleshi, and A. Vancura. Protein acetylation and acetyl coenzyme a metabolism in budding yeast. *Eukaryot Cell*, 13(12):1472–83, 2014.

[157] Mikael Turunen, Jerker Olsson, and Gustav Dallner. Metabolism and function of coenzyme q. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1660(1):171–199, 2004.

[158] D. R. Appling. Compartmentation of folate-mediated one-carbon metabolism in eukaryotes. *Faseb j*, 5(12):2645–51, 1991.

[159] S. Hohmann and P. A. Meacock. Thiamin metabolism and thiamin diphosphate-dependent enzymes in the yeast saccharomyces cerevisiae: genetic regulation. *Biochim Biophys Acta*, 1385(2):201–19, 1998.

[160] R. A. John. Pyridoxal phosphate-dependent enzymes. *Biochim Biophys Acta*, 1248(2):81–96, 1995.

[161] A. Solmonson and R. J. DeBerardinis. Lipoic acid metabolism and mitochondrial redox regulation. *J Biol Chem*, 293(20):7522–7530, 2018.

[162] R. Rodriguez Melendez. [importance of biotin metabolism]. *Rev Invest Clin*, 52(2):194–9, 2000.

[163] T. L. Poulos. Heme enzyme structure and function. *Chem Rev*, 114(7):3919–62, 2014.

[164] Edward V. Quadros. Advances in the understanding of cobalamin assimilation and metabolism. *British journal of haematology*, 148(2):195–204, 2010.

Appendix A

Constructing the biomass objective function

The biomass objective function is an abstractive reaction of GSMs that contain the necessary biomolecular components to generate a unit of cellular dry weight [67]. By performing appropriate stoichiometric weightings, the flux through this reaction directly correspond to the specific growth rate of the organism [68]. In a similar approach to that of [134], the biomass reaction was expanded by adding five preceding reactions, one for each of the following biomass classes: DNA, RNA, proteins and amino acids, carbohydrates and lipids. The components of these reactions were stoichiometrically weighted using appropriate correction factors, such that each generic biomass product had a normalized biomass weight of 1 gDW. Scaling these generic components in a final biomass reaction using experimental weight fractions (i.e. g component gDW^{-1}) will subsequently enable direct predictions of specific growth rates. All calculations are available in the Supplementary Material 'Biomass composition.xlsx'.

DNA

The relative molar abundance of each nucleotide was assumed to be the same as that of *S. limacinum*. To simulate the energetic requirements of DNA polymerization, the corresponding dNTPs were used as reactants, resulting in a reaction on the form

$$c_1\mathrm{dATP} + c_2\mathrm{dGTP} + c_3\mathrm{dCTP} + c_4\mathrm{dTTP} \longrightarrow \mathrm{DNA} + c_5\mathrm{PPi}.$$

The stoichiometric coefficients c_1 - c_5 were calculated by using the molar masses of the respective dNTPs minus the molar mass of the generated pyrophosphate to scale and normalize the molar fractions of each constituent deoxynucleotide (Table A in S1 - Biomass calculations).

RNA

An identical approach was carried out for the RNA composition, which in a similar way was assumed to be the same as that of the template organism (Table A in S1 - Biomass calculations).

Protein

Protein content, as well as the amino acid composition was assumed to be the same as that of the template organism. To account for the energetic demands of ribosomal protein synthesis, one molar equivalent of ATP and two of GTP were added to the reaction [134]. Additionally, charged tRNA molecules were utilized as reactants to include the preceding energy requirements of the ATP-driven synthesis of aminoacyl-tRNA. Molar ratios of

each aminoacyl-tRNA were calculated based on mass percentages of each amino acid, and subsequently scaled in a similar fashion to that of DNA and RNA using a correction factor (Table B1 in S1 - Biomass calculations).

Lipids

The lipid class contents was partly assumed to be the same as that of iCY1170_DHA, with a few exceptions. Mono- and diacylglycerols were removed as available experimental data were not available for these lipid classes, while free fatty acids were added as a separate component. The weight fractions of each constituent lipid class were normalized, and subsequently converted to molar ratios using the corresponding molar masses. The molar masses of the lipid classes containing fatty acyl chains (e.g. triacylglycerol) were determined using the generic fatty acid distribution obtained from FAME analysis of *A.* sp. T66 (Table 3.2). The molar ratios were then scaled by a correction factor, culminating in the set of stoichiometric coefficients used in the lipid biomass reaction (Table B1 in S1 - Biomass calculations).

Carbohydrates

The carbohydrate content was assumed to be the same as that of *S. limacinum*, consisting of the the two enantiomers D- and L-galactose. To estimate the energy requirements for the polymerization of these monomeric carbohydrates, GDP-L-galactose and UDP-D-galactose were employed. Experimental molar ratios were converted to mass ratios, and subsequently scaled by an appropriate correction factor to obtain the stoichiometric coefficients (Table C in S1 - Biomass calculations).

Biomass reaction

The products of these biomass reactions were then added to a final reaction which combines these generic classes at appropriate ratios based on experimental mass fractions

$$s_1\mathrm{DNA} + s_2\mathrm{RNA} + s_3\mathrm{Protein} + s_4\mathrm{Lipids} + s_5\mathrm{Carbohydrates} \longrightarrow \mathrm{Biomass}.$$

Here, the stoichiometric coefficients s_1 - s_5 are simply the normalized mass fraction of each of the major biomass components (Table D in S1 - Biomass calculations). Consequently, the molar weight of the generated biomass is 1 gDW mmol^{-1}, enabling the prediction of specific growth rates (h^{-1}). An auxiliary demand reaction for the biomass metabolite was also added to allow the reaction to carry a non-zero flux.

Coenzymes and cofactors

A sixth biomass reaction was added when performing the gene essentiality analysis to account for the indispensable requirement for coenzymes and cofactors for cellular growth (see Table 5.1). The stoichiometric weights of the cofactors and coenzymes were set at an arbitrarily low level of 10^{-6}. Additionally, inorganic ions were also added based on that of the biomass reaction of iCY1170_DHA. These ions were; iron, potassium, sodium, calcium, copper, magnesium, silicon and boron.

The decisions regarding the intracellular localization of the various coenzymes and cofactors of the biomass reaction were primarily determined based on biochemical information of the role and subcellular localization of their associated enzymes. However, when certain coenzymes were needed in a particular subcellular compartment but no available

Table 5.1: Listing of essential cofactors and coenzymes added to the biomass reaction to increase the scope and validity of the gene essentiality predictions. The various compartments are denoted by the following abbreviations: cytoplasm (c), mitochondria (m) and peroxisome (x).

Metabolite	Compartments	Description
NAD+	c, m, x	Electron carrier used in a wide range of biological redox reactions, needed for maintaining internal redox states and other biological processes in all included intracellular compartments of the model [151].
NADP+	c, m, x	Similar function to that of NAD+, but also important for the regeneration of glutathione [152], as well as being an important cofactor in a wide range of anabolic pathways [151].
FAD	c, m, x	Essential prosthetic group of flavoproteins [153]. Although predominately present in the mitochondrial compartment [154], it is also needed for the activity of cytosolic and peroxisomal flavoproteins (e.g. peroxisomal acyl-CoA oxidase [155]).
Coenzyme A	c, m, x	Central role in cellular metabolism, as well as post-translational and allosteric regulation [156].
Ubiquinone-9	m	Electron-carrier in the respiratory chain [157].
Glutathione	c, m	Antioxidant protecting the cell from oxidative stress [152].
Tetrahydrofolate	c, m	Central role in the one-carbon metabolism [158].
Thiamine diphosphate	c, m	Coenzyme of enzymes that transfer two-carbon units [159].
Riboflavin-5-phosphate (FMN)	c, x	Along with FAD, an important prosthetic group of flavoproteins [153].
Pyridoxal 5'-phosphate	c	Needed for the catalysis of wide range of biochemical reactions (e.g. transamination) [160].
Lipoic acid	m	Essential cofactor of the E2 subunit of various oxoacid dehydrogenase complexes, as well as the H protein of the glycine cleavage system [161].
Biotin	c, m	Required for the transfer of carbon dioxide in carboxylase reactions [162].
Heme b	c, m	Metalloporphyrin bound to hemoproteins of diverse biological functionalities [163].
Adenosylcobalamin	m	Necessary for the activity of the mitochondrial methylmalonyl-CoA mutase [164].
Methylcobalamin	c	Needed for the activity of the cytosolic methionine synthase [164].

intracellular transporter was identified, the cytosolic form of the coenzyme was chosen by default.

For example, while the majority of intracellular riboflavin-5-phosphate (FMN) acts as redox cofactors of mitochondrial dehydrogenases [153], no transporter or mitochondrial source of FMN were identified in *A.* sp. T66. Consequently, a decision was made not to include a gap filling reaction to enable a mitochondrial uptake of this coenzyme, and rather just use the cytosolic version in the biomass reaction.

Appendix B
Calculations of specific uptake rates

The specific uptake rates q_s of the two carbon sources d-glucose and glycerol, as well as the nitrogen source ammonium q_N, were estimated based on the measured substrate concentrations. The conversion to flux rates in mmol gDW^{-1} h^{-1} were performed by dividing this value by the molar mass in g $mmol^{-1}$. As an example, consider glucose with a molar mass of 180.156 g mol^{-1}, and a measured specific uptake rate of 0.259 g gDW^{-1} h^{-1}. The corresponding uptake flux is given as

$$\frac{0.259 \text{ g gDW}^{-1} \text{ h}^{-1}}{180.156 \text{ g mol}^{-1}/1000 \text{ mmol mol}^{-1}} = 1.44 \text{ mmol gDW}^{-1} \text{ h}^{-1}.$$

The uptake fluxes of the remaining substrates were calculated in a similar fashion (Table 5.2).

Table 5.2: Experimental substrate uptake rates, q_s (g substrate gDW^{-1} h^{-1}), and corresponding uptake fluxes (mmol gDW^{-1} h^{-1}) used to constrain the growth predictions of the reconstructed GSM. Also given are the measured ammonium uptake rates (q_N).

Medium	q_s	Uptake flux	q_N	Uptake flux, ammonium
Minimal glucose	0.259	1.44	0.020	1.11
Minimal glycerol	0.225	2.44	0.018	0.99

Appendix C

Pathway map of the steroid biosynthetic pathway

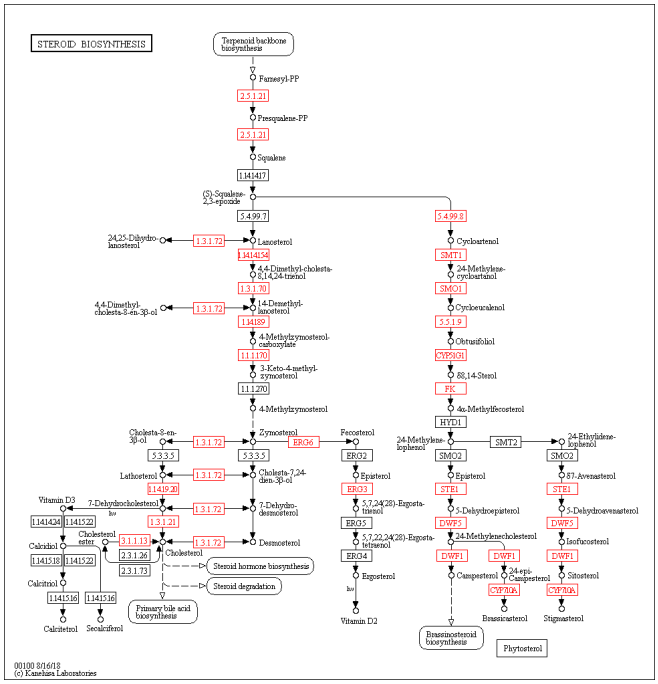


Figure 5.1: Metabolic pathway map of steroid biosynthesis from the KEGG PATHWAY database annotated with the putative metabolic capabilities of *A.* sp. T66. These types of color coded metabolic maps were extensively used in the gap filling procedures during the initial draft model refinement, as well as during subsequent rounds of model curations.

Appendix D
PKS pathway map

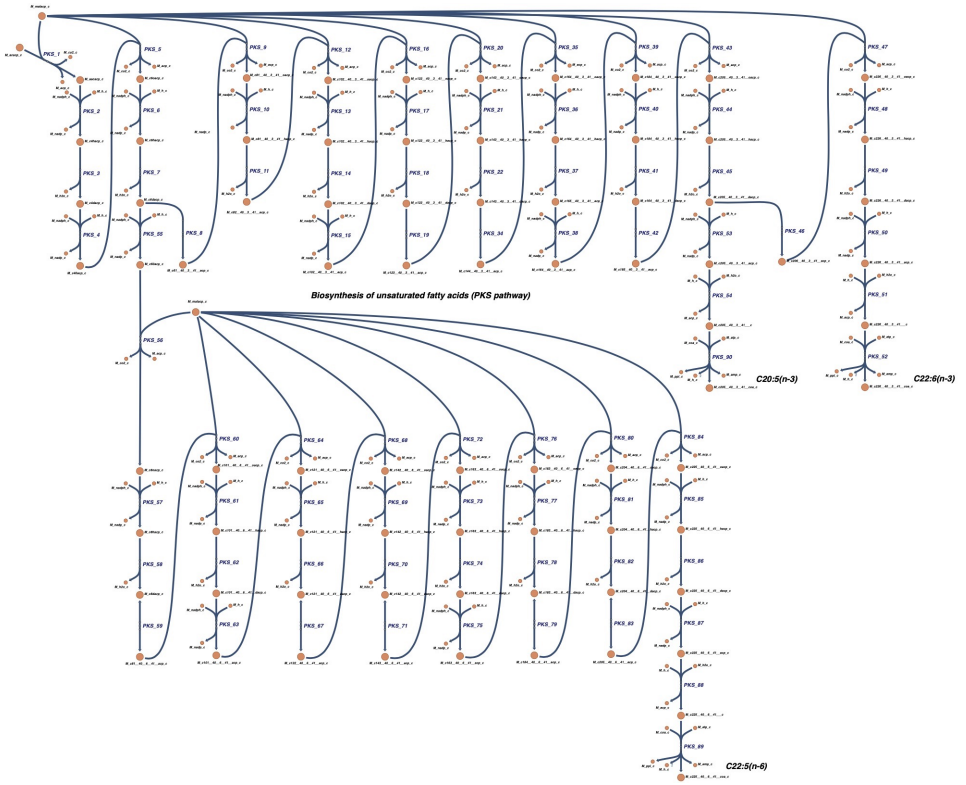


Figure 5.2: Metabolic map of the PKS pathway responsible for the biosynthesis of PUFAs in *A.* sp. T66. The pathway map was generated in a semi-automated fashion using Escher [96]

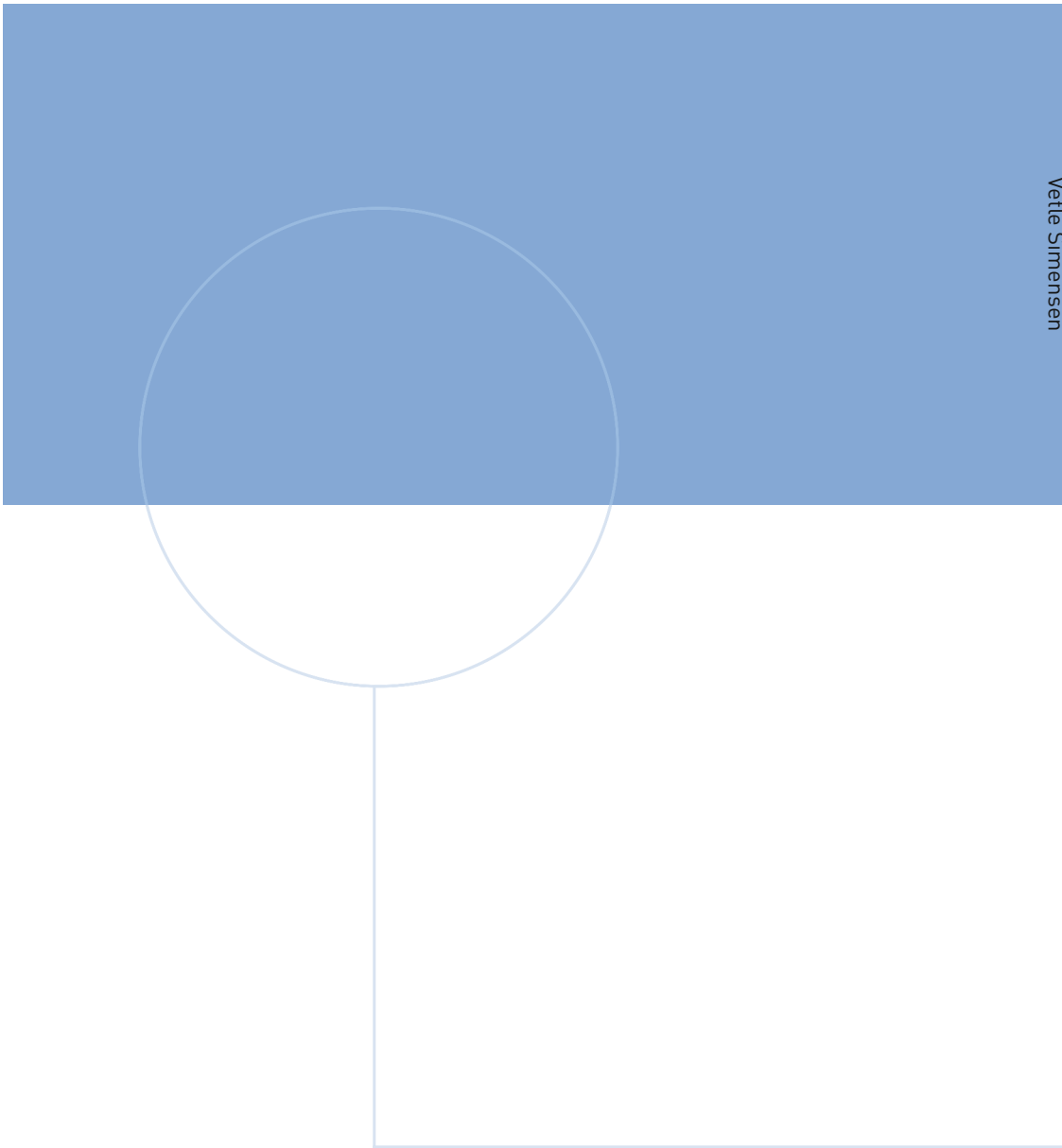
Appendix E

Minimal medium used during gene essentiality predictions

Updated minimal medium used for single gene essentiality analysis, with the addition of essential cofactors which *A.* sp. T66 are unable to synthesize *de novo*. The composition of this medium is taken from [90], except for the carbon source which was changed to that of glucose.

Table 5.3: Updated carbon-limited minimal medium used during the model employment. The growth-limiting uptake rate (mmol gDW^{-1} h^{-1}) of glucose was determined experimentally. The co-factors thiamine and cyanocob(III)alamin were added as *A.* sp. T66 are unable to synthesize these *de novo* [90].

Metabolite	Uptake rate
D-glucose	1.44
Ammonium	10.000
Phosphate	1000
Sulfate	1000
Proton	1000
Water	1000
Oxygen	1000
Calcium	0.010
Boron	0.010
Magnesium	0.010
Silisium	0.010
Copper	0.010
Potassium	0.010
Sodium	0.010
Iron	0.010
Thiamine	1000
Cyanocob(III)alamin	1000



Vetle Simensen

NTNU
Kunnskap for en bedre verden