Development and Analysis of Genome-Scale Metabolic Reconstruction of *Microchloropsis gaditana CCMP526*

Submitted in partial fulfilment of the requirements
of the degree of

Doctor of Philosophy

of the
Indian Institute of Technology Bombay, India
and
Monash University, Australia

by
Akhila George

Supervisors:

Prof. Pramod Wangikar (IIT Bombay)

Prof. John Beardall (Monash University)

Prof. Madhu Chetty (Monash University)

Dr. Santanu Dasgupta (Reliance Industries Ltd)

2018

Dedicated to

my beloved grandfather

J. S. Olickan

Approval Sheet

The thesis entitled "**Development and Analysis of Genome-scale Metabolic Reconstruction of** *Microchloropsis gaditana CCMP526*" by is approved for the degree of **Doctor of Philosophy**
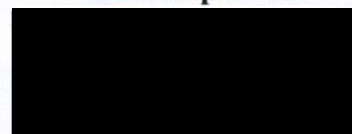
(Prof. Sudip Kundu)
External Examiner

(Prof. Yogendra Shastri)
Internal Examiner

(Prof. Pramod Wangikar)
IITB Supervisor

(Prof. John Beardall)
Monash Supervisor

(Prof. Avinash V. Mahajan)
Chairman

Date: 30th April 2019
Place: Mumbai

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Notice 1

Under the Copyright Act 1968, this thesis must be used only under the normal conditions of scholarly fair dealing. In particular no results or conclusions should be extracted from it, nor should it be copied or closely paraphrased in whole or in part without the written consent of the author. Proper written acknowledgement should be made for any assistance obtained from this thesis.

Notice 2

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Student Name: Akhila George

IITB ID: 124024003

Monash ID: 24731080

Abstract

A eukaryotic microalga, *Microchloropsis gaditana CCMP526,* which belongs to the Class Eustigmatophyceae, produces high amount of lipid that makes it a good potential source of biofuel in algal biotechnology. Manipulation of this alga for enhanced productivity of lipid is well-appreciated but requires a better understanding of its metabolic functioning. Improved understanding of metabolism and metabolic flux control can be achieved through the analysis of a metabolic model constructed for this alga. The systematic approach of metabolic engineering of the alga involves physiological studies of the microalga, generation of a genome-scale metabolic network, flux balance analysis of the metabolic model and identification of the key regulatory pathways.

The curated genome based pathway database for *M. gaditana CCMP526,* MgdCyc, that currently involves 141 metabolic pathways with 1163 metabolic reactions, associated genes, enzymes and metabolites, was developed from the organism-specific genome annotation and data sources such as the KEGG database and the Metacyc database. MgdCyc with its Pathway Tools graphical interface facilitates visualization of pathways and genes in Genome Browser and visualization of functional genomic datasets.

The curated genome-scale metabolic model for *M. gaditana CCMP526,* constituting 720 reactions, was developed to represent the primary metabolism of the alga. The metabolic model was analysed using flux balance analysis to predict the flux of intracellular metabolites in the metabolic network of the alga under different trophic conditions such as phototrophic, heterotrophic and mixotrophic conditions. The metabolic model was validated by comparing the predicted values of the specific growth rate, photosynthetic coefficient and ratio of cyclic electron flow and linear electron flow involved in photosynthesis with their experimental values.

Physiological studies of *Microchloropsis gaditana CCMP526* were conducted to obtain the experimental values of parameters such as growth rate, fatty acid profile, sugar and protein content that were used to validate the genome scale metabolic model of the alga. Studies on growth of the alga in different media with different inoculum concentrations and antibiotic sensitivity were conducted as a preliminary step towards genetic transformation of the alga.

Table of Contents

List of Tables

List of Figures

ABBREVIATIONS

Acronym	Definition
ATP	Adenosine-triphosphate
BLAST	Basic Local Alignment Search Tool
COBRA	Constraint Based Reconstruction Analysis
DEM	Dead end metabolite
DNA	Deoxy ribonucleic acids
EPA	Eicosapentaenoic Acid
FBA	Flux Balance Analysis
GPR	Gene-Protein-Reaction
KEGG	Kyoto Encyclopaedia Of Genes And Genomes
MFA	Metabolic Flux Analysis
MNR	Metabolic Network Reconstruction
PGDB	Pathway/Genome Database
PHF	Pathway Hole Filler
PUFA	Polyunsaturated fatty acids
PPP	Pentose Phosphate Pathway
RNA	Ribonucleic acids
SBML	Systems Biology Markup Language
TCA	Tricarboxylic acid

Chapter 1

Introduction

The unicellular eukaryotic microalga, *Microchloropsis gaditana CCMP526* (previously known as *Nannochloropsis gaditana)* belongs to the Class Eustigmatophyceae that includes fresh water, marine and terrestrial microalgae (Fawley, Jameson, & Fawley, 2015). The alga is an oleaginous marine microalga that has the potential to be commercially used as a lipid producer in the production of biodiesel and as feedstock in mariculture. Therefore, there is renewed interest in metabolic engineering of *Microchloropsis* sp. to increase its growth and lipid productivity.

Traditionally the targets for metabolic engineering have been selected on the basis of the literature or intuitive engineering based on specialised metabolic knowledge. In such cases, the gene predictions turned out to be suboptimal due to the complexity of the metabolic network caused by the regulatory processes at various levels of cellular functioning (Oberhardt, Palsson, & Papin, 2009) . Therefore, the first step in the systematic approach towards metabolic engineering of *Microchloropsis gaditana CCMP526* is the development of its biochemical pathway database and a metabolic model that represents the primary metabolism in the alga.

The genome based pathway database is an overview of the metabolic pathways (at least the central metabolic pathways namely the biosynthetic pathways of lipid, amino acids, nucleotides, pigments and sugars) in the target organism with the corresponding reactions, the associated enzymes and genes. The pathway database helps to understand the metabolism and study the presence of specific variant pathways in the alga. It can be improved with further advancement in genome sequencing level and annotation and it can be used to generate a template of genome scale metabolic model. The analysis of metabolic model predicts the physiological behaviour of algae under different trophic conditions and helps to identify the key regulatory reactions in metabolic functions.

Based on genome annotation, literature analysis, physiological studies and gene expression analysis, a pathway database for *M. gaditana CCMP526* can be developed. A flux-consistent metabolic model can be derived from the pathway database followed by curation. Analysis of the metabolic model predicts the metabolic flux patterns in the alga cultivated under

different trophic conditions such as photoautotrophic, heterotrophic and mixotrophic condition. The developed metabolic model can be used to predict genes for gene knock-out studies. The analysis of the metabolic model can help to find the trophic conditions for better growth or for the better production of target metabolites.

Organization of Thesis

This report is focused on carrying out the initial steps for metabolic engineering of the alga. This involves primarily development of a curated pathway database for *M. gaditana CCMP526* and a validated genome scale metabolic model for the alga and its analysis that will help to design approaches for metabolic engineering of the alga. This report is organized into seven chapters, including this chapter.

This first chapter defines the topic and explains the scope of development of a metabolic model. The second chapter presents a critical assessment of related work reported in the literature. It presents studies of metabolic models available for other alga, analysis and their conclusions. It also points out gaps in literature that makes the current work relevant and the objectives of this project are defined. The third chapter details the methodologies involved in experiments and computational studies on development of the pathway database and genome scale metabolic model, its analysis and validation of the model. In chapters 4 to 6, the results of the work are presented and discussed. In the last chapter, the conclusions drawn from this work are presented together with a section outlining the future scope of the work.

Chapter 2

Literature Review

2.1 Microalgae

Microalgae constitute a group of unicellular or multi-cellular photosynthetic organisms, which inhabit diverse fresh water, marine and terrestrial environments. It is a polyphyletic group that involves taxa in five out of eight eukaryotic groups. Algae evolved through a primary endosymbiotic event in which a cyanobacterium (a photosynthetic prokaryote) was engulfed by an aerobic eukaryote to form an organelle named the plastid. Subsequently, a photosynthetic alga was engulfed by an unrelated heterotrophic eukaryote to form a plastid in them. The primary and multiple secondary and tertiary endosymbiotic events brought novel combinations of genomes, and thus led to a broad phylogenetic distribution in algae (Cock & Coelho, 2011).

Microalgae have a simple cell structure which allows them to grow efficiently at a higher growth rate with an increased photosynthetic efficiency as compared to vascular plants. Some of them are potentially known for their suitability for bio-fuel production in an economically cost-effective and eco-friendly manner (Lee, Chou, Ham, Lee, & Keasling, 2008). Among eukaryotic unicellular microalgae, *Chlamydomonas reinhardtii* (Dal'Molin et al., 2011; Liu, Vieler, Li, Jones, & Benning, 2013; May, Christian, Kempa, & Walther, 2009), *Ostreococcus tauri, Phaeodactylum tricornutum* (Radakovits, Eduafo, & Posewitz, 2011) and *Thalassosira pseudonona* (Armbrust et al., 2004; Hockin, Mock, Mulholland, Kopriva, & Malin, 2012; Jiang, Yoshida, & Quigg, 2012) have been well studied. However, these microalgae do not normally exhibit higher lipid content as compared to other naturally lipid-producing micro algal species. They require extensive genetic modification to enhance their lipid content beyond what that can be achieved through manipulating their cultivation conditions. More importantly, a comprehensive understanding of biosynthetic metabolic pathways related to lipid production is needed to engineer such organisms for biofuel production. Therefore, selection of microalgae that natively express high lipid content and have a completely sequenced genome and genome annotation are good target candidates for metabolic engineering (Radakovits et al., 2012; Thiele & Palsson, 2010).

3

2.1.1 Eustigmatophyceae

The Eustigmatophyceae are a class of stramenopiles that have evolved through multiple secondary or tertiary endosymbiotic events (Cock & Coelho, 2011). This class was drawn from the algal class of Xanthophyceae based on the morphological unique features (Hibberd & Leedale, 1970). The species of *Nannochloropsis* belong to the family Monodopsidaceae from the class of Eustigmatophyceae (Hibberd, 1981). Members of this genus are found mostly in marine environments but can occur in fresh and brackish water. Recently, a new genus of *Microchlorospsis* that constitutes *M. salina* (previously known as *N. salina*) and *M. gaditana* (previously known as *N. gaditana*) was drawn from the class of Eustigmatophyceae (Fawley et al., 2015). Therefore, reports related to *Nannochloropsis* sp. that were published before 2015, are applicable to *Microchloropisis* sp. as well.

Some of the Eustigmatophyceae, such as *N. oculata* (Converti, Casazza, Ortiz, Perego, & Del Borghi, 2009)*, M. salina* (Boussiba, Vonshak, Cohen, Avissar, & Richmond, 1987) and *M. gaditana* (Simionato et al., 2011) have been reported to have high lipid producing characteristics, high biomass accumulation rates and an ability to tolerate a wide range of pH, temperature and salinity. Therefore, they were seen as potential candidates for the commercial production of biofuels. The growth of *Nannochloropsis* sp. has been tested and proven in large scale cultivation such as outdoor ponds (Boussiba et al., 1985, 1987). They can be used as feed stocks in marine aquaculture for mass production especially to cultivate rotifers due to the high fatty acid content of this marine alga (Hirayama, Maruyama, & Maeda, 1989; Koven et al., 1990). The high unsaturated fatty acid content of *Nannochloropsis* sp. makes it highly favourable for larval nutrition (James & Al-Khars, 1990). Due to the high content of eicosapentaenoic acid of *Nannochloropsis* sp.*,* it can also be used as source of EPA in human diet and it was reported to be effective in reducing cholesterol levels (Werman, Sukenik, & Mokady, 2003). Nutrient profiles of EPA-enriched alga were studied and research has been carried out on optimization of EPA production from *Microchloropsis gaditana* (Mitra, Patidar, George, Shah, & Mishra, 2015; Mitra, Patidar, & Mishra, 2015)*.* Studies were conducted to determine the efficiency of *Nannochloropsis* sp. as a potential supplement to chicken's diet to produce poultry products with ω3 fatty acids (Nitsan, Mokady, & Sukenik, 1999). The potential use of *Nannochloropsis* sp. as feedstock in aquaculture, biofuel production and human diet supplements, encouraged research

to study its physiology and conduct optimization studies and genetic engineering to obtain desired products from the algae.

Much research has been conducted to investigate the effect of various cultivation conditions on the growth and biomass composition of microalgae that belong to *Nannochloropsis* species. The environment variables under investigation include light, temperature, pH, sources of carbon and nitrogen and their availability. It was observed that the amino acid composition of *Nannochloropsis* sp. did not vary with changes in light intensity, temperature and nitrogen availability though there was variation in protein content of the cell (Sukenik, Zmora, & Carmeli, 1993).

The growth of *Nannochloropsis* sp. was reported to be light saturated at 200 μmol quanta m^{-2} s^{-1} (Sukenik, 1991). A decrease in cellular content of carbohydrate and lipid with an increase in chlorophyll content was observed for *Nannochloropsis* sp. cultivated under low light, i.e.~ 35 μmol quanta m^{-2} s^{-1} (Renaud et al., 1991; Sukenik et al., 1993). However, an increase in the ratio of content of unsaturated fatty acids to that of saturated fatty acids in *Nannochlorospis* was observed while reducing the light intensity for their cultivation (Renaud et al., 1991). A similar increase was observed in the profile of relative distribution of eicosapentaenoic acid (20:5(n-3)) with a decreased light intensity for cultivation (Sukenik, 1991). The optimum temperature for maximum growth rate of *Nannochlorospis* sp. was found to be 25^{0}C (Sukenik, 1991). Low temperature for cultivation of *Nannochloropsis* sp. increases the cellular content of ω3 poly unsaturated fatty acids in the algae, but this condition discourages growth and thus overall it does not affect the EPA productivity significantly (Sukenik, 1991).

Generally, in microalgae, the storage compounds of lipid and carbohydrate can be accumulated with a decrease in protein content during nitrogen starvation. Similarly, an increase in total carbohydrates and a decrease in protein content and total carotenoids were observed in marine *Nannochloropsis* species (Sukenik et al., 1993). However, the amino acid composition does not vary significantly in *Nannochloropsis* sp. in spite of variations in protein content of the algae with the change in nitrogen availability (Sukenik et al., 1993). Considering the potential of the microalgae as biofuel producers and feedstocks in aquaculture, fatty acid content and composition of the microalgae is important. *Nannochloropsis* sp. can accumulate high amounts of lipid during the stationary phase of growth. Some investigators have concluded that the higher

lipid synthesis in *Nannochloropsis* sp. during stationary phase of growth was due to nitrogen starvation (Suen, Hubbard, Holzer, & Tornabene, 1987). Under nitrogen limitation, the cells fix carbon in presence of light and the carbon tends to flow through biosynthesis pathways to storage molecules such as fatty acids and carbohydrate rather than to protein (Sukenik, 1991). However, there have been other reports which state that lipid accumulation is not due to nitrogen starvation but is due to microelement deficiency that occurs at higher pH (Ben-Amotz, 1985; Boussiba et al., 1987).

The growth rate of *Nannochloropsis* sp. seemed to be adversely affected by the low initial concentration of nitrogen source in the medium, i.e. < 3 mM nitrogen (Sukenik, 1991). It was observed that though the lipid content of *Nannochloropsis* sp. was higher, the lipid production rate was significantly lower in nitrogen-starved culture due to lower growth rate (Boussiba et al., 1985). As the availability of the nitrogen source is reduced, the cellular content of triacylglycerol increases and this in turn affects the relative abundance of fatty acids in the algae (Sukenik, 1991). Triacylglycerol, a storage compound containing neutral lipid that constitutes saturated fatty acids, can be used to produce energy for cellular maintenance during energy shortage (dark period of light-dark cycle) (Sukenik, 1991). It was observed that the cellular content of eicosapentaenoic acid (20:5(n-3)) (EPA), found in a galactolipid, remain consistent even during nitrogen starvation (Sukenik, 1991). This was due to the fact that such polyunsaturated fatty acids primarily form structural lipids like galactolipids which were not affected by such environmental variations. Nitrogen availability can be considered as an operational parameter in the commercial production of *Nannochloropsis* sp. with the desired biochemical composition.

As the biochemical composition of the microalgae changes with the variation in environmental conditions, many studies have been carried out to produce the desired biomass composition by varying the cultivation conditions. The environmental conditions were optimized to increase the content of eicosapentaenoic acid (EPA) in marine *Nannochloropsis* sp. at the cost of the productivity of EPA, since those conditions affected the growth rate adversely (Sukenik, 1991). The cultivation conditions to increase EPA content of the marine alga were nutrient sufficient medium, low light and a lower temperature compared to that optimized for maximum growth rate (Sukenik, 1991).

Polyunsaturated fatty acids (PUFA) are essential for the growth of marine invertebrates and they also play a major role in human diet. Though a significant increase in lipid content in *Nannochloropsis* sp. was observed under nitrogen limitation, the relative distribution of polyunsaturated fatty acids (PUFA) decreased in the alga (Sukenik et al., 1993). Therefore, Sukenik et al. (1993) suggested nutrient sufficient condition with high light intensity at temperature optimal for growth, to cultivate marine *Nannochloropsis* sp. with high PUFA productivity for application in aquaculture (Sukenik et al., 1993). It was reported that an increase in the ratio of content of unsaturated fatty acids (especially C16 and C18 fatty acids) to that of saturated fatty acids in *N. oculata* can be achieved by low light intensity, i.e. ≤ 490 E m^{-2} s^{-1} (Renaud et al., 1991). In the optimization using biochemical engineering that involve optimization of cultivation medium and conditions, there has been a compromise between the growth rate and the cellular content of specific component (Sukenik, 1991).

2.1.2 *Microchloropsis gaditana*

Microchloropsis gaditana sp. which was previously known as *Nannochloropsis gaditana,* was isolated and was found to have features different from other species of *Nannochloropsis* such as *Nannochloropsis oculata* and *Microchloropsis salina* in 1982 (Lubian, 1982). Recently, *Microchloropsis* species have attracted considerable attention by biofuel researchers, as they exhibit high photoautotrophic biomass accumulation rates with a capability to be scaled up to high volume cultures and have a high lipid producing ability.

M. gaditana CCMP526 was isolated from Lagune de Oualidia, Morocco by Billard in 1985. This alga was reported to be a good choice of a model organism for metabolic engineering for biofuel production, as it exhibits higher lipid content with a potential to grow at cell densities greater than 10 gL^{-1} (Jinkerson, Radakovits, & Posewitz, 2013). A feature of this organism is that it can accumulate a large quantity of lipid in the form of triacylglycerides (TAG) during the logarithmic growth stage (Jinkerson et al., 2013). A genetic transformation method involving electroporation at high field strength is established for *Microchloropsis gaditana CCMP526* (Radakovits et al., 2012) . Experimental findings suggest that the *Nannochloropsis* sp. genome is haploid, which enables the organism to exhibit the phenotype instantly after dominant and recessive gene mutations (Jinkerson et al., 2013; Kilian, Benemann, Niyogi, & Vick, 2011;

Radakovits et al., 2012; Weeks, 2011). Therefore, a systematic genome-scale metabolic model approach can be used to engineer this type of organism in order to achieve high lipid production at commercial level as high productivity and high lipid yield of the alga can be achieved by metabolic engineering.

2.2 Genome-scale metabolic reconstruction

Metabolic networks are cascades of enzyme-enzyme relations. It is considered complete only if the enzymes are linked according to its biological context and tagged with its genomic information (Reed et al., 2006). To understand the whole functioning of an organism, a comprehensive reconstructed metabolic network is always required. Metabolic network reconstruction (MNR) provides necessary biological information to develop suitable microbial metabolic models for predicting the cellular phenotype of an organism. MNR is performed through assigning Gene-Protein-Reaction (GPR) associations to the annotated genes from the sequenced genome. It integrates the genome annotation data, metabolic biochemistry and physiology of an organism to annotate the GPR associations. In addition, it tags the active metabolic pathways that are linked to GPR associations.

The real value and challenge of a reconstructed network lies in accomplishing a successful annotation. A high-quality of genome annotation only can assist in understanding the metabolic process of an organism (Stein, 2001). Further, a meaningful annotation only can guide to investigate the possible and impossible biochemical reactions that are present in the metabolic network of an organism. Thus, a reconstructed network includes the number of genes, proteins (i.e., enzymes catalysing metabolic reactions), metabolites and reactions that take part in the metabolic activity of an organism, wherein they are categorized, interconnected and represented in a network fashion (Feist, Herrgård, Thiele, Reed, & Palsson, 2009). Furthermore, genome-scale reconstruction directs hypothesis-driven discovery, by integrating high throughput data with biological discovery process and bridging the genotype-phenotype gaps in unicellular or multi-cellular microbial systems.

However, obtaining a high-quality MNR is challenging nowadays as automated reconstructed network are prone to errors, which are caused by incomplete annotations and poor

sequence similarities. Consequently, the automated MNR are subjected to an additional step of refinement process, where the gaps and the errors in the MNR are fixed and validated through exploring various functional prediction computational tools, literature based context and integrating high-throughput experimental 'omics' datasets, such as genomics, proteomics and metabolimics (Saha, Chowdhury, & Maranas, 2014). The metabolic networks thus completely re-constructed with no gaps are typically used to develop genome-scale metabolic models for deducing the metabolic potential and fitness of an organism.

2.2.1 Development of Metabolic reconstruction

The metabolic network reconstruction process consists of four major stages. The four stages are creating a draft reconstruction, manual reconstruction refinement, conversion of a network into a mathematical model and network evaluation. (Thiele & Palsson, 2010)

The draft reconstruction is a collection of genome-encoded metabolic functions which depends mainly on the genome annotation available. The generation of draft reconstruction involves obtaining genome annotation, identifying candidate metabolic functions, obtaining candidate metabolic reactions for these functions, assemble draft reconstruction. The first stage of creating a draft reconstruction for *M. gaditana* was thus to obtain the genome annotation (mainly the genes and their proteins with the database links and evidences). The accuracy of metabolic pathways in the draft reconstruction strongly depends on the accuracy of genome annotation used for the reconstruction. The more the gene annotations available, the more reliable is the metabolic reconstruction. So it is important to collect the most recent version of the genome annotation. (Thiele & Palsson, 2010)

In the next stage, i.e. manual reconstruction refinement, the draft construction is re-evaluated, curated and refined. The metabolic reactions and their Gene-Protein-Reaction (GPR) associations in the metabolic pathways are individually considered for organism-specific literature and refinement. The manual re-evaluation is important because not all the annotations have a high confidence score and the databases referred to are organism-unspecific so consequently the reactions that are not there in the target organism may be included in the pathways predicted (Thiele & Palsson, 2010).

In the third stage of metabolic network reconstruction, the reconstruction is converted into mathematical format and condition-specific models are defined. The final metabolic model is obtained after multiple iterations of validation and refinement. The fourth stage involves the verification, re-evaluation and validation of the metabolic model.

2.2.2 Pathway Tools software

Using genome annotation, the draft reconstruction can be generated and refined using the Pathway Tools software. This software infers metabolic pathways by analyzing the metabolic annotations with respect to a reference database of metabolic pathways such as Metacyc. The pathway prediction is based on a computed score that reflects the likelihood that the pathway is present. The score value depends on the total number of reactions in the pathway, the number of reactions for which the annotated genes of the target organism are available and the number of pathways in which the same enzyme is involved. (Thiele & Palsson, 2010)

Pathway Tools software can also be used to infer which genes are likely to code for the missing enzymes in the metabolic pathways. It identifies the transport proteins in the genome and infers transport reactions from the free text transport function descriptions that are present in the genome. It facilitates the editing of information regarding genes, proteins, reactions, pathways, and chemical compounds (Peter D Karp et al., 2010).

2.2.3 Metabolic Flux Analysis

Metabolic flux analysis (MFA) quantifies the flow of materials in metabolism resulting in generation of flux maps, which in turn help to revise the metabolic model developed. There are several approaches for MFA which can be selected on the basis of network size, whether the biological system can be evaluated at the steady state, and the availability of details of reactions. The flux analysis using FBA that give insights regarding the key regulatory pathways involved in the lipid biosynthesis, are selected based on the data available and the feasibility. This basically helps to identify the critical branch points and reactions of which the manipulations may produce a significant effect on lipid content of microalgae (Allen, Libourel, & Shacharâ€¢ Hill, 2009; Thiele & Palsson, 2010) .

Flux balance analysis defines the metabolic network as a linear programming optimization problem. The constraints are imposed by the steady state mass conservation of metabolites in the system. The intracellular flux of the metabolites and the flux ratios of the pathways can be estimated computationally using flux balance analysis (FBA). The flux and its ratios can be obtained at branching points in the pathways using the tool such as COBRA (Thiele & Palsson, 2010) .

2.2.4 Metabolic reconstruction of algae

Constructing genome-scale metabolic models assists in contextualization of high-throughput data by imposing constraints on genome scale reconstructed metabolic networks based on known experimental measurements. Further, it provides complete insights on critical pathways by investigating the flux distributions in the metabolic network. Besides, metabolic modelling helps to study multi-cellular systems by comparing the metabolic models of the species with their phenotypic differences and the analysis. It also enables the discovery of network properties, which includes the existence of loops, optimal pathway usage, pathway redundancy and metabolite connectivity (Oberhardt et al., 2009).

However, during construction of metabolic models, the occurrence of mis-annotation of genes, the existence of inactive isozymes and pleiotropy reduces the reliability of the predictions that are obtained from the constructed model. In some cases, uncharacterized enzyme kinetics, complexity in network interaction, and unexpected regulation may demand a dynamic modelling process such as ensemble modelling for specific pathways to improve predictions. As the common targets for genetic manipulation are regulatory genes, the lack of characterization of regulatory networks makes it unreliable to predict the metabolic engineering targets. The metabolic network is a simplified representation of cellular function and the very simplification may make the metabolic reconstruction challenging to involve novel phenomena in the model. Thus, the application of modelling is limited to analysis and refinement of the knowledge of already characterized cellular systems. In the metabolic modelling of eukaryotic microalgae, the major challenge is the lack of data regarding the compartmentalization and the transport mechanisms between compartments (Contador, Rizk, Asenjo, & Liao, 2009; Dal'Molin et al., 2011; Oberhardt et al., 2009) .

Cristiana et al. (2011) has reported a genome-based metabolic network model named AlgaGEM for *Chlamydomonas reinhardtii*, which covered the metabolism of a compartmentalized algae cell. This model considered the compartments as cytoplasm, mitochondrion, plastid and microbodies. It included functions of 866 unique open reading frames, 1862 metabolites, 2249 gene-enzyme-reaction association entries and 1725 unique reactions. The model also agreed with the simulation of growth and algal metabolic functions obtained from the literature. Likewise, a genome-scale metabolic model called 'AraGEM' for *Arabidopsis thaliana* represented the primary metabolism of a compartmentalized plant cell. This model considered 1419 unique open reading frames, 1748 metabolites, 5253 gene- enzyme-reaction association entries and 1567 unique reactions (de Oliveira Dal'Molin, Quek, Palfreyman, Brumbley, & Nielsen, 2010). The model was validated through the simulation of plant metabolic functions based on literature context. Moreover, the compartments that included in AraGEM model was primarily based on the literature search and the current databases that related to *Arabidopsis thaliana.* The compartmentalization was manually performed through considering the knowledge of organelle functions and the localization of some isozymes.

In 2010, the Dal'Molin research group reported a C4 genome-scale model to investigate flux distribution in mesophyll cells and bundle sheath cells during C4 photosynthesis in C4 grasses, such as maize, sugarcane and sorghum (Dal'Molin, Quek, Palfreyman, Brumbley, & Nielsen, 2010). This model seems to be the first large scale genomic model, which involved metabolic interactions between two cell types, that is, M and BS cells. Katsunori et al.(2011) developed a genome-scale metabolic model for the cyanobacterium *Synechocystis* sp. PCC6803 and this was validated using ^{13}C metabolic flux analysis (Dal'Molin et al., 2011; Yoshikawa et al., 2011) .

In *Chlamydomonas reinhardtii,* AraGEM based compartmentalization data was considered for constructing the genome-scale metabolic model. However, as the transport reactions between the cytoplasm and organelles or extracellular spaces were poorly annotated, some reactions were added manually based on literature. Likewise, the compartmentalization data of *Chlamydomonas reinhardtii* can be used to include compartmentalization in the metabolic modelling of *Microchloropsis* (Dal'Molin et al., 2011; de Oliveira Dal'Molin et al., 2010) .

2.3 Gaps in Literature

Microchloropsis gaditana seems to be a promising renewable feedstock for biofuel production. The annotated genome and the genetic transformation method for *M. gaditana* are reported in literature. This may help to rapidly implement the engineering techniques to improve the organism into a high lipid producing strain (Jinkerson et al., 2013; Radakovits et al., 2012; Yee, Ahmad, & Cha, 2012).

The main challenge in the genetic modification of *Microchloropsis gaditana* is to find key regulatory reactions that can be manipulated by metabolic engineering to produce a desirable strain. Biochemical pathway database with genes, enzymes and reactions that represent metabolism in an organism can facilitate the systematic analysis of cellular functions in the organism. Development of a pathway database for the alga can contribute towards the generation of a metabolic model and the pathway database can be further improved with advancements in genome sequencing and genome annotation. There are accessible biochemical pathway databases available for bacteria (P. D. Karp, 2002), protozoa (Doyle et al., 2009), green algae (May et al., 2009), plants (Mueller, Zhang, & Rhee, 2003; Urbanczyk-Wochniak & Sumner, 2007; Van Moerkercke et al., 2013) and trees (P. Zhang et al., 2010). As there is no biochemical pathway database developed for *M. gaditana CCMP526,* the work of developing a pathway database included in the thesis can contribute towards the further advancement of genome annotation and analysis of metabolic capability of the alga.

The development of a genome-scale metabolic model facilitates the identification of key regulatory pathways and strategies to reroute the metabolite so that a higher amount of the desirable product is produced. Then the study of the regulatory networks, along with the genome-scale metabolic model, helps to suggest the possible gene manipulations necessary to improve the characteristics of the strain (Dal'Molin et al., 2011; de Oliveira Dal'Molin et al., 2010; Jinkerson et al., 2013; Radakovits et al., 2012; Vieler et al., 2012). Though, there is metabolic model available for *Microchloropsis sp.* (Loira et al., 2017; Shah, Ahmad, Srivastava, & Jaffar Ali, 2017)*,* there is no metabolic model developed for *M. gaditana CCMP526.* There were physiological studies reported for *M. gaditana CCMP526.* However, the biomass composition such as sugar content, protein content and the fatty acid profile, required for the

formulation of biomass objective function in the flux balance analysis of the metabolic model were not reported in literature.

2.4 Objective

The overall objective of this thesis was to develop a validated metabolic network reconstruction as the first step towards metabolic engineering of *M. gaditana CCMP526.*

Objective 1: Construction of a pathway database for *M. gaditana CCMP526*

The first step towards metabolic engineering of *Microchloropsis gaditana CCMP526* is the development of its biochemical pathway database. The pathway database helps to understand the metabolism and study the presence of specific variant pathway in the alga. Unlike metabolic model, the pathway database can afford to involve flux-inconsistent pathways that can be further investigated with experiments. Moreover, the pathway database can be investigated for improvement or metabolic modelling of the alga using a user-friendly web interface. Our objective was to construct a pathway database from genome information and literature that represents the metabolism of *Microchloropsis* sp. as completely as possible. Development of the pathway database is based on genome annotation, literature, physiological studies and gene expression analysis. The genome annotation, based on which pathway database is generated, is very limited. Therefore, extensive curation of the pathway database is required to make it as complete as possible.

Objective 2: Development of metabolic model and analysis

This objective was to construct a genome-scale metabolic model which is an overview of metabolic pathways (at least the central metabolic pathways, biosynthesis pathways of lipid, amino acids, nucleotides, pigments and sugars) in the target organism with the corresponding reactions, the associated enzymes and genes. The metabolic network can be curated to make the metabolic model flux consistent. The metabolic model can be analysed using flux balance analysis of the metabolic model to predict the metabolic fluxes in the alga under different trophic

conditions such as phototrophic, heterotrophic and mixotrophic conditions. The analysis also predicts the growth rate of the alga under different trophic conditions. The predictions from analysis of the metabolic model can be compared with experimental values and the metabolic model can be validated.

Objective 3: Physiological study of *M. gaditana CCMP526*

Physiological studies of the alga can be used for validation of developed metabolic model. The objective was to study the physiology and the biomass composition of the target organism from the literature and experiments. The biomass composition of the alga is required to formulate the biomass formation equation. In addition, the physiological studies allow comparison of metabolic model predictions with the experimental values and thus support the reconstruction and refinement of the metabolic model. These studies involve monitoring growth, estimation of macromolecular biomass composition and preliminary studies required for genetic transformation of *M. gaditana*.

Chapter 3

Materials and Methods

3.1 Development of MgdCyc, a biochemical pathway database for *M. gaditana CCMP526*

3.1.1 Development of initial build

Genome sequence and annotation of genome assembly version 1.2 of *Microchloropsis gaditana CCMP526* submitted by the Colorado School of Mines (Radakovits et al., 2012), was obtained from The National Centre of Biotechnology (https://www.ncbi.nlm.nih.gov/). The details of the genome annotation of assembly v.1.2 which was used as input file for creating pathway database are given in Table 12 (Appendix). The pathway database of *M. gaditana CCMP526* was constructed from genome annotation in Genbank file format and the genome sequence in FASTA format using the pathologic module of Pathway Tools software v20.0 (P. D. Karp, Paley, & Romero, 2002). Taxonomic pruning of reactions to reduce false positive pathway predictions was enabled during the automated build of the database. The inclusion of pathways in the network depends on the pathway score which is based on genome annotation and nature of reactions. Pathway score indicates the likelihood that the pathway is present in this alga. Pathway prediction score cut-off for the automated build was set to the default value of 0.15 without compromising specificity and sensitivity of pathway inclusion in pathway database. The default value was reported to give the best trade-off between sensitivity and specificity where the value of pathway prediction score cut-off ranges from zero to one.

Pathway prediction algorithm

The pathologic component of Pathway Tools software infers reactions present in organism based on the enzymes in the annotated genome. It can also predict the metabolic pathways present in the organism based on the inferred reactions and other factors such as the expected taxonomic range of the pathway and the presence of GPR associated key reactions in the pathway. This can

16

be carried out by computing the pathway score for all pathways in MetaCyc, a multiorganism database of metabolic pathways and enzymes that are curated from scientific literature (Caspi et al., 2006). Pathway score of a pathway gives a measure of the likelihood that the pathway is present in the subject organism.(Peter D Karp, Latendresse, & Caspi, 2011)

Pathway score (PS) is calculated as follows:

$$PS = \frac{\sum_{r \in R} RS(r)}{|R|} + T$$

Where $RS(r)$ is reaction score of all enzyme-catalysed reactions r in the pathways and |R| is the number of reactions in the pathway. T is given a value higher than zero if the subject organism is in the taxonomic range of the pathway. The reaction score is computed as follows:

$$RS = P + U + K$$

Where P is the presence score that carries a value of 0.2 if an enzyme catalysing reaction is present in the subject organism and otherwise the value becomes zero. U is the uniqueness score that carries a value between 0.6 where the reaction is present in a single pathway and 0 where the reaction is present in a large number of Metacyc pathways. K is given a value of 0.5 if the reaction is the key reaction in the pathway. The pathologic component includes the pathway if its pathway score is higher than the pathway prediction score cut-off.

3.1.2 Curation of pathway database

The pathway database for *M. gaditana CCMP526* was curated based on the genome annotation, literature, physiology of the alga and homology search analysis. The steps involved in curation of initial build are given in Figure 1.

Figure 1: Steps in curation of pathway database, MgdCyc

3.1.2.1 Assigning probable metabolic enzymes

Some enzymes that were not recognized by automated name matching procedure were curated using the Pathologic component of Pathway Tools software version 20.0 (Peter D Karp et al., 2010), that involves creating additional enzyme-to-reaction assignments. This involves assigning

reactions to the probable enzymes found by the software by referring KEGG, Metacyc and other reference databases.

3.1.2.2 Identification of missing enzyme using Bayesian method

The reactions that are not associated with GPR (gene-protein-reaction) can be assigned to corresponding annotated genes by using the Pathway Hole Filler (PHF). The PHF, a component of Pathway Tools software, is used to find the pathway holes that are reactions for which the genome of target organism lacks enzymes (Green & Karp, 2004). The program identifies and re-evaluates the candidate sequences in the genome based on evidence from a homology search (such as E-value, alignment length, the rank of the candidate in the BLAST output), from the pathway context of the missing reactions and operon-based data. The BLAST of the target genome was carried out against the collection of protein sequences from entries from GenPept, Swissprot, PDB, PRF, PIR and NCBI Reference Sequence (RefSeq) project.

The PHF runs in two phases: training and prediction. The initial phase involves training of the Bayes classifier with the known reactions in pathway-genome database (PGDB) to predict which candidates have the desired function and which do not. The program will then make predictions for pathway holes in the PGDB of interest. Those pathway holes could be manually filled based on the values of probability. Probability was calculated based on obtained suggestions based on evidence from a homology search (such as E-value, alignment length, the rank of the candidate in the BLAST output), from the pathway context of the missing reaction, operon-based data and the requirement of the associated pathway in the model (Green & Karp, 2004). Candidates predicted with probability threshold of 0.9 were reported to be 71% precise (Green & Karp, 2004).

3.1.2.3 Curation using FBA

Most of the gaps in the metabolic network that makes the metabolic model flux inconsistent can be filled by Metaflux. Metaflux is a multiple gap filling method, component of Pathway Tools software version 20.0. General development mode of metaflux develops a feasible model that can generate non-zero fluxes for some reactions given a biomass reaction, nutrients and secretion. Metaflux carries out flux balance analysis considering the try sets and numerical value

parameters, called weights, provided by the user. The trial set is a set of candidates to be considered for filling the incomplete model to get a non-zero flux through the network. The weight associated with every candidate is added to the global objective function and MILP maximizes that objective. A positive weight indicates the need to include the associated candidates to the metabolic model of interest.

The given biomass reactions included (7Z)-hexadecenoyl-CoA, (9Z,12Z)-hexadeca-9,12-dienoyl-CoA, alpha-linolenoyl-CoA, oleoyl-CoA, myristoyl-CoA, linoleoyl-CoA, linoleoyl-CoA, icosapentaenoyl-CoA, stearoyl-CoA, palmitoyl-CoA, palmitoleoyl-CoA, gamma-linolenoyl-CoA, arachidonoyl-CoA, 4-hydroxy-l-proline, pentadecanoyl-CoA, di-homo-γ-linolenate, ATP, CTP, UTP, GTP, dATP, dCTP, dGTP, dTTP, chlorophyll-a, D-ribopyranose, GDP-L-fucose, L-alanine, L-arabinopyranose, UDP-L-rhamnose, , UDP-alpha-D-galactose, UDP-alpha-D-glucose, alpha-D-xylopyranose, L-arginine, L-aspartate, L-cysteine, L-glutamate, L-histidine, L-isoleucine, glycine, L-leucine, L-lysine, L-methionine, L-ornithine, L-phenylalanine, L-proline, L-serine, L-threonine, L-tryptophan, L-tyrosine and L-valine. The general development mode was run with nutrients set to be glucose, nitrate, sulphate, phosphate, protons, magnesium, water, oxygen and bicarbonate. The secretions were set to be phosphate, protons, water, oxygen and carbon dioxide.

The cost for adding one reaction outside the taxonomic range of the PGDB from MetaCyc to the model was set to be -200. The weight for adding the reverse of an irreversible reaction from the PGDB to the model was -100. The weight for adding the reverse of an irreversible reaction from MetaCyc to the model was -200. The weight for adding one reaction within the taxonomic range of the PGDB from MetaCyc to the model was -40. The weight for adding one reaction from MetaCyc with an unknown taxonomic range was -80. The weight for adding a spontaneous reaction was -1. The weight for adding a transport reaction from MetaCyc to the model was -300.

3.1.2.4 Curation based on homology search analysis

Homology sequences of genes of enzymes involved in pathways such as the TCA (tricarboxylic acid) cycle, photosynthesis, the pentose phosphate pathway and amino acid (alanine, arginine, cysteine, glutamate, glutamine, proline, pyrrolysine) biosynthesis pathways in phylogenetically closely related algae such as *M. gaditana B-31* are searched against the genome assembly of *M.*

gaditana CCMP526 using BLASTN 2.7.0+ (Z. Zhang, Schwartz, Wagner, & Miller, 2000). BLAST (Basic Local Alignment Search Tool), is a heuristic algorithm to find the similarity in primary sequences such as nucleotide and protein sequence. A function was assigned to nucleotide sequence based on the E-value, query coverage, identity score and bit score, resulting from BLAST analysis.

3.2 Development and analysis of metabolic model for *M. gaditana CCMP526*

The PGDB (Pathway/Genome database), MgdCyc was exported in SBML format from Pathway Tools software to the Matlab environment. The SBML file was then converted to mathematical model and carried out FBA using COBRA toolbox (Hyduke et al., 2011). The biomass formulation equation was formed using experimental from literature. The experimental values were taken from graphs using WebPlotDigitizer (Drevon, Fursa, & Malcolm, 2017). The development of genome-scale metabolic model from the pathway database is outlined in Figure 2.
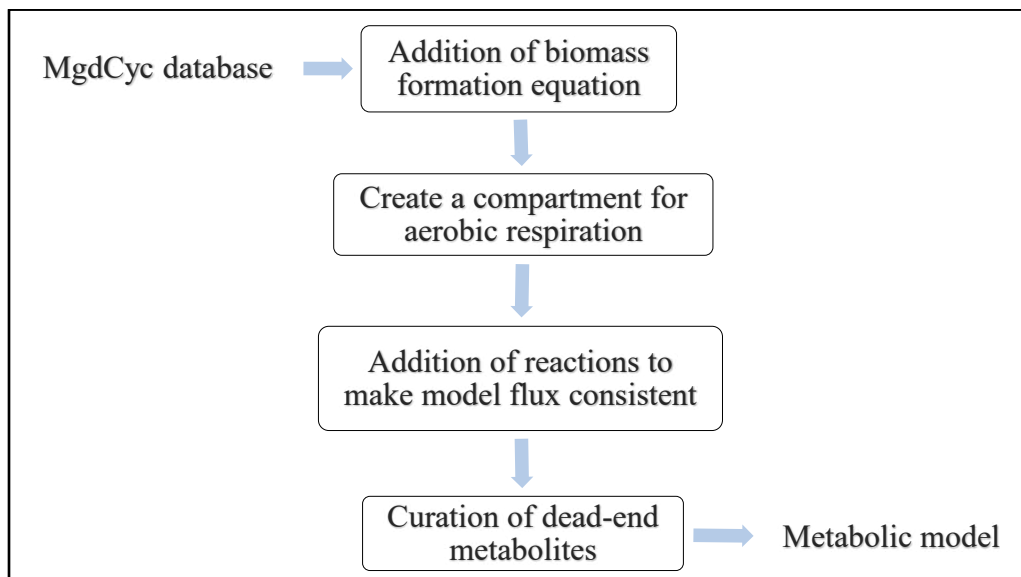
Figure 2: Steps in development of metabolic model for *M. gaditana*

3.2.1 Curation and refinement

The curation of reactions in metabolic network involved correction of directionality and addition of annotations by reviewing literature and the database such as KEGG (Kyoto Encyclopaedia of Genes and Genomes), Biocyc, and metabolic models of other microalgae.

A dead end metabolite (DEM) is a metabolite that lacks requisite metabolic or transport reaction for its production or consumption in the metabolic network (Mackie, Keseler, Nolan, Karp, & Paulsen, 2013). The analysis of the DEMs was carried out by understanding the biochemistry and metabolic context of the DEM and by reviewing the literature and databases such as Metacyc (curated database of metabolic pathways) and Biocyc (pathway/genome database). The dead end metabolites were analysed and removed, if required, by removing or adding metabolic reactions or transport reactions associated with DEM to the metabolic network. The gaps associated with dead end metabolites were analysed and any of the following actions was taken.

- If the reaction (associated with DEM) is a general reaction that associates with general terms like protein, lipid, sugar, then the reaction is removed since the network refers the metabolites more specifically.
- If the reaction is isolated and plays no significant role in the physiological behaviour of the organism, the reaction is removed from the draft network, thus the dead end metabolite.
- If a relevant reaction is isolated and there is no upstream reaction, gene annotation, presence of this reaction in phylogenetically closely related microalgae was referred to add upstream reaction to the draft network.

If necessary, pathways associated with dead end metabolites are included in the network if any of the following was satisfied.

- If the presence of the pathway in *Microchloropsis* species is supported by literature.
- If other relevant or unique reactions in the pathway are associated with annotated gene.

3.2.2 Flux Simulation

Constraint-based reconstruction and analysis (COBRA) uses various linear, quadratic, mixed integer linear quadratic and nonlinear optimization programming solvers (Schellenberger et al., 2011) to quantify and predict the cellular metabolism and phenotype of an organism. It integrates the physiochemical and biochemical information from the MNR, and with implied biological constraints it predicts the possible phenotypic states for the metabolic network, under the given set of conditions. COBRA performs multiple tasks, such as FBA analysis, flux variability analysis (FVA), MOMA, gene deletion studies, gap filling and visualization of flux distributions (Schellenberger et al., 2011).

3.2.3 Linear programming formulation

A system of linear algebraic equations is formed from the steady state material balance of all metabolites in metabolic network reconstruction. The linear equations are represented in the form of

$$S * v = 0$$

where S is the stoichiometric matrix that includes the coefficient of the metabolites in the reaction equation and v is the flux vector that includes objective function. Since the algal system is assumed to be at the steady state, the stoichiometric matric is the steady state mass balances on metabolites and the flux vector is zero except the reaction which is the objective function. The stoichiometric matrix is in the form of $m * n$ where m represents metabolites and n represents reactions. (Shastri & Morgan, 2005; Stephanopoulos, Aristidou, & Nielsen, 1999)

Maximize biomass subject to:

$$\sum_j s_{ij} v_j = 0 \; for \; every \; i \in M_i$$

$$v_{min} \; \leq \; v_i \; \leq \; v_{max}$$

where s_{ij} is the stoichiometric coefficient of the ith metabolite in the jth reaction and v_j is the flux of the jth reaction. The flux can take values from a given range, which is generally 0 to 1000 for intracellular metabolites and -1000 to 1000 for reversible reactions.

The above-mentioned constraints were applied to the model and flux balance analysis of metabolic model was carried out.

3.3 Physiological studies of *M. gaditana CCMP526*

Physiological studies were carried out to help the metabolic network reconstruction of *M. gaditana CCMP526* and the validation of the metabolic model. The physiological aspects that were explored involve growth, macromolecular composition and photosynthesis. Preliminary study for genetic transformation of *M. gaditana* is also conducted that may help metabolic engineering of this strain.

3.3.1 Algal strain and Culturing

The marine microalga, *Microchloropsis gaditana CCMP526* was isolated from 32.8333°N 9°W Lagune de Oualidia, Morocco and the axenic culture is preserved in National Center for Marine Algae and Microbiota (NCMA) in Maine, USA. The microalga was cryopreserved using 6% dimethyl sulfoxide (Cañavate & Lubián, 1997; Cañavate & Lubiń, 1995). The microalga was precultured in f/2 medium (Guillard, 1975; H. Ryther & Guillard, 1962) with pH 8.0 in 100mL Erlenmeyer flask under continuous light of 74 µmol/m^2/s with constant orbital shaking (110 rpm) at 24ºC. The alga was subcultured every two weeks after checking for bacterial or fungal contamination by streaking on LB plates.

3.3.2 Growth study

The growth of *M. gaditana CCMP526* was monitored by counting cells using a hemocytometer or by measuring absorbance at 680nm and 750nm using a UV/visible light spectrophotometer. These measurements can be used to calculate specific growth rate, doubling time, divisions per day and maximum biomass productivity of alga (Neidhardt, Ingraham, & Schaechter., 1990).

Specific growth rate (μ) in h^{-1} can be calculated by the following equation where x_t is the absorbance of culture at time t.

$$\mu = \frac{(\ln x_2 - \ln x_1)}{(t_2 - t_1)}$$

The doubling time (t_d), division per day and maximum biomass productivity can be calculated using the following equations where m_t is the biomass at the time of t.

$$t_d = \frac{(\ln 2)}{\mu}$$

$$Divisions \quad per \quad day = \frac{1}{(t_d)}$$

$$Maximum \quad Biomass \quad Productivity = \frac{(m_2 - m_1)}{(t_2 - t_1)}$$

3.3.2.1 Selection of media for cultivation

M. gaditana CCMP526 was cultivated in 100 mL media in 500mL Erlenmeyer flask under continuous light of 74 μmol/m^2/s with constant orbital shaking (110 rpm) at 24°C. The growth was monitored and growth parameters were calculated. Since the algal cells started turning pale from 4th day of the growth onwards, we attempted to prevent the culture from turning pale by maintaining pH of the media and using high concentration of nitrate in media. The following media were used to find the effect of Tris-Cl buffer (pH 8) and high concentration of nitrate (17 mM) on growth of *M. gaditana*. Note that f/2 medium normally has 0.88 mM sodium nitrate.

- f/2 medium with 40 mM Tris Cl buffer and 17mM sodium nitrate
- f/2 medium with 17mM sodium nitrate
- f/2 medium with 40 mM Tris Cl buffer
- f/2 medium

3.3.2.2 Selection of different nitrate concentration

M. gaditana CCMP526 was cultivated in 100 mL f/2 medium in 500mL Erlenmeyer flask under continuous light of 74 μmol/m^2/s with constant orbital shaking (110 rpm) at 24°C. The culture was cultivated in media with different nitrate concentrations of 0.05 mM, 0.1 mM, 0.3 mM, 0.9 mM and 17mM. The growth in each medium was monitored and growth parameters were calculated.

3.3.3 Sugar estimation

A sample (1 mL) of the alga was harvested during steady state growth under sterile conditions and centrifuged at 4,500 rpm for 5 minutes. The supernatant was discarded and the pellet was washed with 0.5 M ammonium bicarbonate to remove medium constituents. The pellet was resuspended in 0.5M ammonium bicarbonate and taken for estimation of sugar content using the phenol sulfuric acid method (DuBois, Gilles, Hamilton, Rebers, & Smith, 1956; Zhu & Lee, 1997). The standard curve obtained for estimation of sugar is given in Figure 12 (Appendix).

3.3.4 Protein estimation

A sample (10 mL) of the alga was harvested during steady state growth and centrifuged at 4000 rpm for 5 minutes. The supernatant was removed and 1 mL of 1N sodium hydroxide solution was added. The sample was vortexed and incubated at 90°C for 10 minutes. The sample was, then, centrifuged to collect the supernatant and the treatment of pellet 1N sodium hydroxide solution was repeated three times. The supernatant was collected after every centrifugation of 1N sodium hydroxide solution treated sample. The supernatant was then taken for protein estimation using Bradford's protein assay reagent kit. The standard curve obtained for estimation of protein is given in Figure 13 (Appendix).

3.3.5 Dry Biomass estimation

Aliquots of 10 mL algal suspension was filtered onto preweighed glass fiber filters (0.22 μm Millipore) under reduced pressure. The filtered biomass was washed with 20 mL of 0.5M

ammonium bicarbonate to remove medium components from the filter and biomass. The filtered biomass was then dried at 100ºC to a constant weight, cooled down in vacuum desiccator and weighed. (Zhu & Lee, 1997)

3.3.6 Estimation of fatty acid profile

Gas chromatography - Mass Spectrometry (GCMS) was used for the estimation of fatty acids present in *M. gaditana* using Omegawax 250 column (30m * 0.25mm * 0.25µm) that produces reproducible analyses with fatty acid methyl esters. The fatty acid derivatization involved one-step lipid extraction and fatty acid methylation described by Garcés and Mancha (1993) and Jacobsen et al. (2011). The GC method was developed based on the protocol described by de la Vega et al. (2011). (de la Vega, Díaz, Vila, & León, 2011; Garcés & Mancha, 1993; Jacobsen, Rosgaard, Sakuragi, & Frigaard, 2011)

3.3.7 Preliminary studies for genetic transformation of *M. gaditana CCMP526*

The growth characteristics of culture in f/2 medium and TAP medium with inoculum sizes of 1%, 5% and 10% were analysed by measuring absorbance at 750nm using a UV/visible light spectrophotometer . The cell concentrations (cells mL^{-1}) are calculated from optical density of culture at wavelength of 750nm (Lopes & Vasconcelos, 2011). This study gives insights regarding the better medium for the cultivation of culture, the duration of growth phases and the cell density which would help to design the protocol for genetic transformation. The antibiotic sensitivity study of *M. gaditana* was carried out with the antibiotics chloramphenicol, kanamycin and hygromycin with each antibiotic supplied at concentrations of 30µg/mL, 100µg/mL and 300µg/mL. This supports the selection of a vector to be used for genetic transformation of *M. gaditana* in which the selection process of the genetically transformed strains would be based on the resistance of the culture against an antibiotic.

Chapter 4

MgdCyc – a biochemical pathway database for
Microchloropsis gaditana CCMP526

4.1 Introduction

Understanding the metabolism of algae is crucial to metabolic engineering of algae. An organism-specific pathway database that represents at least the reactions and metabolites that are relevant to the targeted physiology helps to understand the metabolism of algae. Such a database can help to construct a metabolic model to predict the metabolic behaviour of the alga in a given set of external conditions or a given genetic perturbation. However, the accuracy of the prediction using a metabolic model depends on how precise the process of development of the metabolic network was. Moreover, the pathway database can be investigated for improvement or metabolic modelling of the alga using a user-friendly web interface.

In *M. gaditana,* many enzymes that are relevant in central carbon metabolism are not assigned a gene in the available genome annotation; however, it is currently needed to understand the metabolism in *Microchloropsis gaditana.* Therefore, the genome annotation available for *M. gaditana* CCMP526 was used to develop the best possible approximation for its integrated pathway-genome database. The pathway database can be used to predict the metabolic composition of *M. gaditana CCMP526.*

Our objective was to construct a pathway database from genome information and literature that represents the metabolism of *Microchloropsis* sp. as complete as possible.

4.2 Initial build

The initial build of MgdCyc, the pathway database for *M. gaditana CCMP526,* involved 65 pathways, 836 enzymatic reactions with 485 enzymes and 814 metabolites, which constitute 11% of the genome annotation. The distribution of pathways (other pathways such as activation/inactivation, tRNA charging, aromatic compounds metabolism, protein modification,

metabolic regulator metabolism and sugar derivative synthesis are not included in the Table 1) in different subsystems in the initial build is shown in Table 1.

Table 1 : Subsystems and pathways in the initial build and MgdCyc database

Subsystems	Number of pathways in initial build	Number of pathways in final version of MgdCyc
Amine and polyamine metabolism	3	6
Amino acid metabolism	12	24
Carbohydrate metabolism	3	6
Cofactors, prosthetic groups, electron carrier metabolism	7	14
Fatty acid and lipid metabolism	9	18
Nucleoside and nucleotide metabolism	14	28
Secondary metabolism	2	4
Inorganic nutrients metabolism	3	6
Generation of precursor metabolites and energy	3	21

It was found that the pathways that are crucial to central carbon metabolism such as the TCA cycle, the Calvin cycle and the pentose phosphate pathway were not present in the initial build of MgdCyc. This is due to the absence of enzymes catalysing the reactions in those pathways in the database. Therefore, extensive curation of the initial build was required to develop a pathway database that represents the primary metabolism of *M. gaditana CCMP526*.

4.3 Curation of pathway database

4.3.1 Assigning probable metabolic enzymes

Out of 220 candidates, 79 probable metabolic enzymes were assigned reactions by manually analysing them by referring to KEGG (Kyoto Encyclopedia of Genes and Genomes) and Metacyc databases. The unassigned probable metabolic enzymes were found to be either not metabolic enzymes or non-specific enzymes or duplicates of other enzymes. There were 26 probable enzymes that were found to be not metabolic enzymes and 65 non-specific enzyme names that cannot be assigned any reaction. Specific metabolic enzymes could not be found for 18 probable enzyme matches. Thus, a total of 188 probable enzymes was analysed and the remainder were found to be duplicates or isozymes of other probable matches. The assigned probable enzyme matches are given in Table 13 (Appendix).

Rescoring pathways using a pathway scoring algorithm, after assigning reactions to probable metabolic enzymes, with pathway prediction score cut-off of 0.15, resulted in the addition of another ten pathways (listed in Table 2) to the pathway database.

Table 2 : Inferred pathways in the pathway database, MgdCyc

Pathway ID in MgdCyc	Pathway name
PWY3DJ-12	Ceramide *de novo* biosynthesis
PWY-4081	Glutathione-peroxide redox reactions
PWY-46	Putrescine biosynthesis III
PWY-5136	Fatty acid & beta oxidation II
PWY-6019	Pseudouridine degradation
PWY-6368	3-phosphoinositide degradation
PWY-6368	3-phosphoinositide degradation
PWY-6599	Guanine and guanosine salvage II
PWY66-21	Ethanol degradation II
PYRUVDEHYD-PWY	Pyruvate decarboxylation to acetyl CoA

4.3.2 Addition of pathways based on literature

There were missing pathways in the initial build that were reported to be present in *M. gaditana.* Based on literature (listed in Table 14 (Appendix) with references), some pathways with their variant pathways were added to the database. Thus pathways such as the tricarboxylic acid (TCA) cycle, the glyoxylate cycle, acetyl-CoA biosynthesis II, β oxidation, biosynthesis of amino acids except that of glycine, aspartate and serine, galactolipid biosynthesis I, ceramide *de novo* biosynthesis, CDP-diacylglycerol biosynthesis I, phosphatidate biosynthesis (yeast), sulfoquinovoysl diacylglycerol biosynthesis, tetrapyrrole biosynthesis, methyl erythritol phosphate pathway, mevalonate pathway, mono trans. poly-cis decaprenyl phosphate biosynthesis, ergosterol biosynthesis, 7-dehydroporiferasterol biosynthesis, plant sterol biosynthesis and cholesterol biosynthesis were added.

4.3.3 Identification of missing enzyme using Bayesian method

The pathway-genome database (PGDB) of *Saccharomyces cerevisiae S288c,* YeastCyc of version 19.5 was used to train the Bayes classifier involved in Pathway Hole filler (PHF) program since its PGDB was built with genome annotations of higher quality. Using PHF, 717 pathway holes (reactions that lack associated enzymes in database) were found in 187 pathways in the database of *M. gaditana* and one or more candidates were found to fill 470 of these holes. Given a probability threshold of 0.9, 29 enzymes were assigned to fill pathway holes in the database (given in Table 15).

The procedure was repeated by training the Bayes classifier with the PGDB of *E.coli,* EcoCyc of version 20.0. Using PHF, the number of pathway holes found in the draft network was 688 in 183 pathways and one or more candidates were found to fill 441 of these holes in the database of *M. gaditana*. 19 pathway holes (given in Table 16) were filled by assigning candidate enzymes that scored probability value above 0.9 to the pathway holes.

In total, 48 enzymes were assigned to reactions in the pathway database of *M. gaditana CCMP526*. The pathway prediction score was again calculated for the base pathways in the database and 55 variant pathways with lower prediction score were removed from the database.

This curation process involving pathway hole filler and removal of low scoring variant pathways resulted in the addition of 76 pathways (given in Table 17) to the pathway database.

Macromolecule reactions were removed from the database because either they are general reactions or not directly relevant to the metabolic behaviour of the alga. Some generic small molecule reactions were also removed from the pathway database.

4.3.4 Gap filling based on physiology

Based on gap filling method using Metaflux, 81 reactions were found to fill the gaps in the pathway database so that the resulting metabolic network can produce biomass in flux balance analysis. Those reactions and corresponding pathways were manually analysed and added to the pathway database. Considering ChlamyCyc, PGDB of *Chlamydomonas reinhardtii* and AraCyc, PGDB of *Arabidopsis thaliana* as reference models, pathways involved in 5-aminoimidazole ribonucleotide biosynthesis, nucleotide biosynthesis, tetrahydrofolate biosynthesis, folate transformation, chlorophyll a biosynthesis, coenzyme A biosynthesis, phosphopantothenate biosynthesis, chorismate biosynthesis, icosapentaenoate biosynthesis, alanine biosynthesis II, gluconeogenesis, nucleotide sugar synthesis and sulfate reduction were added to the pathway database. Ten enzymes corresponding to the pathway holes in some of the pathways added were found using pathway hole filler (listed in Table 18). Then, those pathways were removed from database since their pathway prediction score was lower than the default value of 0.15.

4.3.5 Homology-based gene prediction

There were some pathways and its variants which were reported to be in *M. gaditana* but not present in the database due to pathway prediction score lower than the default threshold value, i.e 0.15. The holes in some of those pathways could not be improved by Bayesian method using Pathway Hole Filler. Therefore, an attempt was carried out to predict gene and enzymes corresponding to reactions involved in such pathways. Homolog genes of corresponding enzyme were searched in the genome of *M. gaditana CCMP526* using BLAST (Basic Local Alignment Tool) for those reactions in the Calvin cycle, the pentose phosphate cycle, glycolysis and the TCA cycle.

The genome annotation of *M. gaditana B-31* (Corteggiani Carpinelli et al., 2014) in which 10695 genes are annotated, is significantly better than the genome annotation of the *CCMP526* strain (Radakovits et al., 2012) in which 3557 genes are annotated. Therefore, the gene sequence of corresponding enzyme in *M. gaditana B-31* was used as query to find a homolog in the genome of the *CCMP526* strain.

The predicted enzymes involved in glycolysis, TCA, calvin cycle and pentose phosphate pathway with the best hits obtained in the genome of *M. gaditana CCMP526* using BLASTn search are given in Table 19, Table 20, Table 21 and Table 22 (Appendix) respectively. The % identity shows the extent to which two sequences have the same residues at the same position at an alignment. The alignment coordinates of the sequence are given by s.start and s.end. Bit score is a log-scaled version of total score, i.e., it gives the magnitude of the search space you would have to look through before you would expect to find a score as good as or better than this one by chance. The expectation value, E-value is the indicator of the validity of match. Smaller the E-value, better the match is (McGinnis & Madden, 2004). The prediction of genes involved in TCA cycle suggests the presence of at least the partial pathway.

Genes corresponding to two enzymes involved in the glycolysis (given in Table 3), five enzymes in the TCA cycle (given in Table 4), two enzymes in the Calvin cycle (given in Table 5) and three enzymes in the pentose phosphate pathway (PPP) (given in Table 6) were found in the genome of *M. gaditana CCMP526*.

Table 3: Predicted genes corresponding to enzymes involved in reactions in glycolysis

Gene ID	Enzyme name	Reaction ID
NGA_A000018	Glucose-6-phosphate isomerase	PGLUCISOM-RXN
NGA_A000019	Enolase	2PGADEHYDRAT-RXN
NGA_A000020		

Table 4: Predicted genes corresponding to enzymes involved in reactions in TCA cycle

Gene ID	Enzyme name	Reaction ID
NGA_A000001	Citrate synthase	CITSYN-RXN

Gene ID	Enzyme name	Reaction ID
NGA_A000002		
NGA_A000003	Isocitrate dehydrogenase	RXN-9951
NGA_A000004	2-oxoglutarate dehydrogenase E1 component	2OXOGLUTARATEDEH-RXN
NGA_A000005		
NGA_A000006	Succinate dehydrogenase iron sulfur protein	SUCCINATE-DEHYDROGENASE-UBIQUINONE-RXN
NGA_A000007	Succinate dehydrogenase flavoprotein subunit	
NGA_A000008	Succinate dehydrogenase subunit 4	
NGA_A000009	Succinate dehydrogenase cytochrome b subunit	
NGA_A000010	Fumarate hydratase	FUMHYDR-RXN
NGA_A000011		

Table 5: Predicted genes corresponding to enzymes involved in reactions in calvin cycle

Gene ID	Enzyme name	Reaction ID
NGA_A000012	Fructose-1,6-bisphosphatase	F16BDEPHOS-RXN
NGA_A000013		
NGA_A000014		
NGA_A000015		
NGA_A000016	Phosphoribulokinase	PHOSPHORIBULOKINASE-RXN
NGA_A000017		

Table 6 : Predicted genes corresponding to enzymes involved in reactions in PPP

Gene ID	Enzyme name	Reaction ID
NGA_A000021	6-phosphogluconolactonase	6PGLUCONOLACT-RXN
NGA_A000022	Transketolase	1TRANSKETO-RXN
NGA_A000023		
NGA_A000024		
NGA_A000025		

NGA_A000026		
NGA_A000027		
NGA_A000028		
NGA_A000029		
NGA_A000030		
NGA_A000031		
NGA_A000032	Transaldolase	TRANSALDOL-RXN
NGA_A000033		
NGA_A000034		
NGA_A000035		
NGA_A000036		

Some reactions that involve tRNA charging and those containing generic terms such as DNA, electron acceptor and protein were excluded from the database since they are not specific reactions. Some of the reversible reactions, especially the ones that involve transfer of phosphate from ATP and quinone were made irreversible by referring to the databases such as MetaCyc and KEGG.

4.4 MgdCyc

The curated pathway database for *M. gaditana CCMP526*, MgdCyc currently features 141 pathways with 1163 reactions with 537 enzymes, 1007 compounds and associated genes. The inclusion of reactions in the database was based on presence of corresponding enzyme in the database or evidence based on literature or corresponding pathway score. This database has 739 dead-end metabolites and 495 reactions with missing enzymes that include generic reactions where macromolecules are involved, translation of mRNA and secondary metabolism. In central metabolism, dead-end metabolites are involved in different pathways such as nucleotide biosynthesis, tetrahydrofolate biosynthesis, folate transformation, chlorophyll *a* biosynthesis, coenzyme A biosynthesis, phosphopantothenate biosynthesis, chorismate biosynthesis, icosapentaenoate biosynthesis, alanine biosynthesis II, gluconeogenesis, nucleotide sugar synthesis and sulfate reduction.

The distribution of pathway in the database to different subsystems is given in Table 1. Pathways in the database are given in Table 23 (Appendix) with their pathway prediction score. The pathway database can be explored using user-friendly Pathway Tools graphical interface and the genes can be visualized in the genome browser available in Pathway Tools.

4.5 Conclusions

MgdCyc provides a curated biochemical pathway database for *M. gaditana CCMP526* that features 141 metabolic pathways with associated genes, enzymes and metabolites. This pathway database provides a user-friendly Pathway Tools graphical interface that supports visualization of functional genomics datasets. MgdCyc facilitates further analysis of metabolism in *M. gaditana CCMP526* and comparative studies of metabolism across different species. The database can be updated as the genome sequencing and genome annotation under genome sequencing project of *M. gaditana CCMP526* progresses. It also provides guidance in predicting and annotating the unknown genes in the alga. Sixty one enzymes that are missing in the available genome annotation were identified in the genome sequence of *M. gaditana CCMP526*. The MgdCyc database can be downloaded as flat file from Biocyc distribution of pathway/genome database or can be accessed with Pathway Tools web.

Chapter 5

Development and analysis of metabolic model for

Microchloropsis gaditana CCMP526

5.1 Introduction

Understanding the metabolic flux control is a key objective of metabolic engineering (Stephanopoulos et al., 1999). Metabolic flux balance analysis of the metabolic model of an organism facilitates the simulation of metabolism in the organism under different environmental conditions and genetic perturbations. This provides information on metabolic fluxes through different pathways and insights regarding metabolic flux control and thus helps metabolic engineering for the enhancement of desired traits of the organism.

5.2 Formulation of Biomass objective function

The formulation of biomass objective function for flux balance analysis is based on biomass composition of the alga that involves biomass precursors of ribonucleic acids (RNA), TCA (DNA), fatty acids, sugar and protein. The coefficients of biomass precursors that form the biomass formation equation were derived from biomass composition of the alga that is given in (Volkman, Brown, Dunstan, & Jeffrey, 1993)

Table 7. The biomass composition, including amino acids, fatty acids and sugars, was taken from that reported for *Microchloropsis salina* (previously known as *Nannochloropsis salina)* by Volkman et.al (1993). A particular fatty acid composition is shown by most of the species of the class, Eustigmatophyceae (Mourente, Lubian, & Odriozola, 1989). The composition of nucleotides was taken from that reported for *Nannochloropsis* sp.(Rebolloso-Fuentes, Navarro-Pérez, García-Camacho, Ramos-Miras, & Guil-Guerrero, 2001). Chlorophyll *a* is present in the alga, but the alga lacks chlorophyll *b* and *c*, therefore composition of only chlorophyll *a* was included in the biomass equation (Owens, Gallagher, & Alberte, 1987). The reported biomass composition and derivation of coefficients to form biomass formation equation for amino acids, fatty acids, sugar, chlorophyll, DNA and RNA are given in Table 24, Table 25, Table 26, Table

27 and Table 28 (Appendix) respectively. An ATP requirement of 36.5 mmol ATP /g biomass was found as growth associated maintenance energy, i.e., the energy required for transport of biomolecules, polymerization of macromolecules and biosynthetic processes in the alga. The non-growth associated maintenance energy that is required for DNA repair, cell wall maintenance and pH control in the alga was assumed to be 1.50 mmol ATP/g biomass, as reported for *C.reinhardtii* by Boyle and Morgan (Boyle & Morgan, 2009). The proportion of the biomass precursors was included in the stoichiometric metabolic model as biomass formation equation. (Volkman, Brown, Dunstan, & Jeffrey, 1993)

Table 7: Coefficients in biomass formation equation

Biomass Components	Precursors	Coefficient (mmol/g dry weight)
Amino acids	Glycine	0.178
	L-alanine	0.192
	L-arginine	0.102
	L-aspartate	0.146
	L- cysteine	0.017
	L-lysine	0.091
	L-leucine	0.125
	L-isoleucine	0.086
	L-glutamate	0.165
	L-histidine	0.032
	L-methionine	0.031
	L-phenylalanine	0.088
	L-proline	0.150
	L-serine	0.120
	L-tyrosine	0.059
	L-tryptophan	0.016
	L-valine	0.120
	L-threonine	0.124

	Hydroxy-L- proline	0.004
	L-ornithine	0.003
	Aminobutyric acid	0.012
Fatty Acids	Myristoyl-CoA	0.025
	Palmitoyl- CoA	0.122
	Palmitoleoyl-CoA	0.141
	(7Z)-hexadecenoyl- CoA	0.0004
	(9Z,12Z)-hexadeca-9,12-dienoyl- CoA	0.0004
	Stearoyl- CoA	0.004
	Oleoyl- CoA	0.033
	Linoleoyl- CoA	0.006
	Gamma-linolenoyl- CoA	0.002
	Di-homo-gamma-linolenoyl CoA	0.003
	Arachidonoyl- CoA	0.015
	Icosapentaenoyl- CoA	0.06
Sugar	UDP-β - L-arabinopyranose	0.011
	GDP-beta-L-fucose	0.073
	UDP-alpha-D-galactose	0.065
	Chrysolaminarin	0.064
	GDP-alpha-D-mannose	0.015
	UDP-beta-L-rhamnose	0.091
	D-ribopyranose	0.035
	UDP-alpha-D-xylose	0.030
Nucleotides	dTTP	0.002
	dATP	0.002
	dGTP	0.003
	dCTP	0.003
	UTP	0.014

	ATP		0.014
	GTP		0.016
	CTP		0.016
Pigment	Chlorophyll a		0.019
Growth maintenance	ATP		36.5
Non growth maintenance	ATP		1.50

5.3 Curation of metabolic model based on FBA

A flux inconsistent metabolic model for *M. gaditana CCMP526* was generated from MgdCyc, the pathway database of *M. gaditana CCMP526,* using Cobra Toolbox in Matlab environment. Photosynthesis light reactions were replaced by two reactions of cyclic and linear electron flow as shown below.

Linear electron flow:

$$8\ Light +\ 2\ NADP + 3\ ADP +\ 3\ Pi\ + 2\ H_2O\ \rightarrow\ 2\ NADPH\ + 3\ ATP\ +\ O_2\ +\ 2\ H^+$$

Cyclic electron flow:

$$2\ Light + 2\ ADP + 2\ Pi\ ->\ 2\ ATP$$
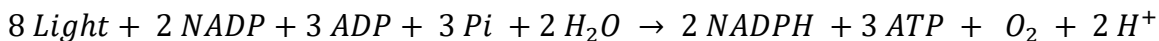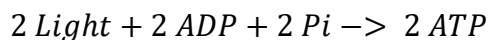
Exchange reactions were added to the metabolic model to facilitate the uptake and/or secretion of 17 extracellular metabolites due to mass-balancing requirement. The electron transfer reactions involved in aerobic phosphorylation and ATP synthesis were assigned compartments of cytosol (denoted by '[c]'), mitochondrial inner membrane (denoted by 'CCO__45__MIT__45__OMEM') and mitochondrial inter membrane space (denoted by '[m]). A reaction to produce chrysolaminarin, the storage sugar that was reported to be present in *M. gaditana*(Wang et al., 2014), was manually added to the metabolic model.

The ability of the metabolic model to produce individual biomass component was tested using Cobra Toolbox in Matlab environment. The metabolic model was unable to produce 13 biomass precursors (out of 49 biomass precursors) under phototrophic condition using flux balance analysis (FBA). Referring to ChlamyCyc (version 5.0), curated pathway database of

Chlamydomonas reinhardtii (Schläpfer et al., 2017) and AraCyc (version 12.0), curated pathway database of *Arabidopsis thaliana*(Mueller et al., 2003), 79 reactions involved in different pathways such as nucleotide biosynthesis, tetrahydrofolate biosynthesis, folate transformation, chlorophyll *a* biosynthesis, coenzyme A biosynthesis, phosphopantothenate biosynthesis, chorismate biosynthesis, icosapentaenoate biosynthesis, alanine biosynthesis II, gluconeogenesis, nucleotide sugar synthesis and sulfate reduction were added to the metabolic network to fill gaps in the metabolic model for modelling purpose. Since the lack of NADPH-glutamate dehydrogenase enzyme activity and presence of glutamine sythetase activity in *Nannochloropsis oculata* were experimentally reported, pathways involving glutamine synthetase and glutamate synthetase were added for nitrate and ammonium assimilation to the metabolic model (Everest, Hipkin, & Syrett, 1986). The metabolic network was thus curated to generate a flux consistent model that produces all the biomass precursors under flux balance analysis, i.e., a metabolic network involving 1211 reactions and 1025 metabolites was made flux consistent.

The metabolic model was further analysed to curate dead-end metabolites and physiologically irrelevant reactions using Cobra toolbox in Matlab. Using this tool, 164 root no production gaps (metabolites that are associated with consuming reactions but no producing reactions) and 205 root no consumption gaps (metabolites that are associated with producing reactions but no consuming reactions) were found in the metabolic model. Most of those metabolites were involved in generic reactions where macromolecules are involved, translation of mRNA and secondary metabolism. Therefore, the reactions that involve those gaps were manually analysed and removed from the metabolic model that represents primary metabolism of *M. gaditana CCMP526.* However, the remaining gaps were filled based on gene annotation and databases of Metacyc and KEGG.

Quality of the metabolic model was ensured by carrying out standard protocols (Thiele & Palsson, 2010). Stoichiometrically balanced cycles or Type III extreme pathways that can carry flux despite closed exchange reactions (Thiele & Palsson, 2010) were not found in the metabolic model. The metabolic model was also checked for ATP production without energy inputs. It was ensured that no ATP was produced under simulation of metabolic under phototrophic condition with unlimited bicarbonate available as input, but no photon uptake.

The metabolic model thus curated features 720 reactions with compartments of cytoplasm, mitochondrial inter membrane space, mitochondrial inner membrane and extra cellular space.

5.4 Metabolic flux topologies under different conditions

The topological properties of the metabolic network were predicted by simulation of metabolic model with flux constraints on few parameters that corresponded to the physiochemical behaviour of the alga. The metabolic model was simulated under different trophic conditions such as photoautotrophic, heterotrophic and mixotrophic conditions since *Microchloropsis* sp. were reported to grow under these conditions (Das, Lei, Aziz, & Obbard, 2011; Fang, Wei, Zhao-Ling, & Fan, 2004). Under photoautotrophic condition, the alga fixes net carbon dioxide/bicarbonate in the presence of light where carbon dioxide/bicarbonate and light act as carbon source and energy source respectively. On the other hand, under heterotrophic condition, an organic substrate such as glucose or ethanol acts as carbon and energy source for growth of the alga. Under mixotrophic condition, the alga utilizes both the organic and inorganic carbon source simultaneously with light for its growth. However, in nature, algae grow under phototrophic condition during the day as it takes sunlight and carbon dioxide for growth, whereas it grows under heterotrophic condition during the night as it degrades the stored energy source into simple carbon molecules and uses them for growth.

5.4.1 Phototrophic simulation

Phototrophic conditions were simulated for the algal metabolic model by a two-step optimization method. In case of *M. gaditana,* the alga seems to have a higher affinity for bicarbonate than for carbon dioxide, therefore inorganic carbon in the form of bicarbonate was used as carbon source and photon (light) as the energy source for flux balance analysis of the metabolic model (Huertas, Espie, Colman, & Lubian, 2000; Huertas & Lubian, 1998; Munoz & Merrett, 1989; Sukenik et al., 1997). The first step was to simulate steady state growth of the alga by maximizing the growth rate under unlimited light and carbon source, while fixing the net photosynthetic rate at 1.55 mmol O_2/g dry weight/h that was reported for *Nannochloropsis* sp.*(CCAP 211/78)*(Raso, van Genugten, Vermuë, & Wijffels, 2012). The predicted maximum

growth rate was 0.031 h^{-1} and that value is found to be close to the growth rate observed in the experiment given in Section 0. The second step was to fix the growth rate at the predicted value and minimize the photon uptake rate to make the system energy efficient. The minimum light uptake rate required for the steady state growth of the alga was found to be 13.4 mmol/g dry weight/h. The predicted carbon uptake rate was found to be 1.0 mmol/g dry weight/h. The major predicted fluxes of intracellular metabolites in the metabolic model are represented in Figure 3. Under photoautotrophic condition, the major metabolic flux was predicted to flow through Calvin cycle and gluconeogenesis.
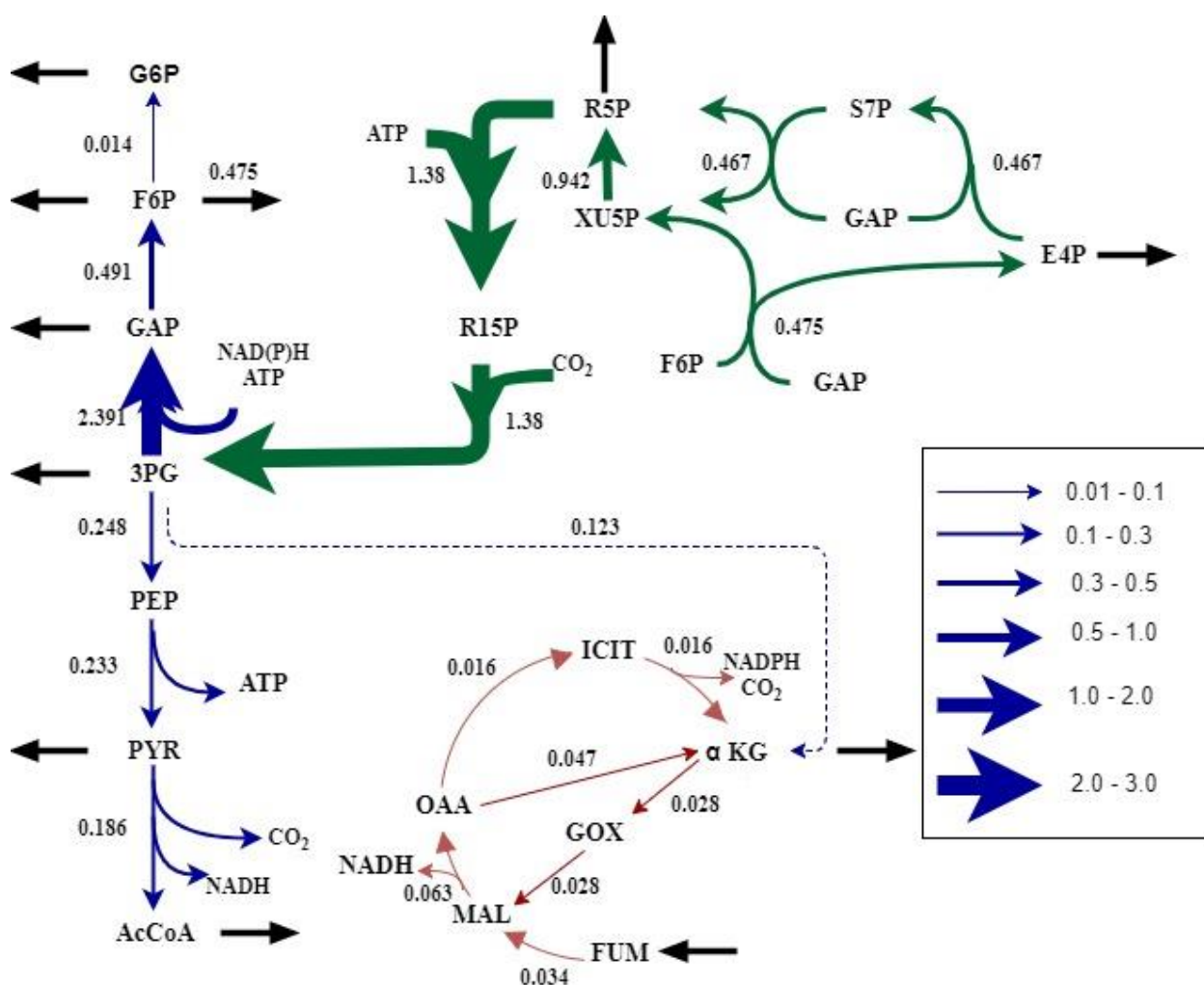

Figure 3: Autotrophic metabolic flux map.

The values of metabolic flux through individual metabolic reactions are mentioned near the arrows and the range of flux values are represented by the thickness of arrows. The reactions in

green, blue and red colour represents reductive pentose phosphate pathway, gluconeogenesis and TCA cycle respectively.

5.4.2 Heterotrophic simulation

Heterotrophic conditions were simulated by maximizing growth rate of the alga in the absence of light while fixing the respiratory rate at a reported experimental value of 0.055 mmol O_2/g dry weight/h for *Nannochloropsis* sp.(Fang et al., 2004). The growth rate was predicted to be 0.005 h^{-1} for the alga under heterotrophic condition which is a much lower value when compared to that under phototrophic conditions. A similar behaviour was observed in experiments reported for *Nannochloropsis* sp. in the literature (Fang et al., 2004; Marudhupandi, Sathishkumar, & Kumar, 2016). This could be due to the simultaneous utilization of the carbon source for biosynthetic process of biomass precursors and energy production. Unlike photoautotrophic conditions, the major metabolic flux flows through pathways of glycolysis and TCA cycle. A significant portion of carbon seems to be lost in the form of carbon dioxide in the TCA cycle that might also contribute to the low growth rate of the alga. The glucose uptake rate of the alga under the simulation of heterotrophic condition was found to be 0.06 mmol/g dry weight/h. The major predicted flux through the algal metabolic network under heterotrophic condition is represented in Figure 4.

5.4.3 Mixotrophic simulation

The mixotrophic conditions were simulated by maximizing growth rate while fixing the maximum uptake rate (i.e. lower bound of exchange reaction) of carbon sources at the values predicted under phototrophic and heterotrophic condition in previous sections, and the oxygen evolution rate at 1.719 mmol O_2/g dry weight/h in the presence of unlimited light. The predicted growth rate of the alga was found to be 0.042 h^{-1} that was higher than that under phototrophic condition. A higher growth rate of *Nannochloropsis* sp. under mixotrophic condition was reported in literature (Cheirsilp & Torpee, 2012; Das et al., 2011; Fang et al., 2004; Xu, Cai, Cong, & Ouyang, 2004). The second step was to minimize the photon uptake rate while fixing the biomass production rate at the predicted value and the oxygen evolution rate at 1.719

mmolO$_2$/g dry weight/h. The minimized photon uptake rate was predicted to be 15.14 mmol/g dry weight/h.

The cyclic electron flow increases under mixotrophic condition to balance the ATP/ NADPH ratio inside the cell. The value of CEF/ (CEF + LEF) increases to 0.065 under mixotrophic condition. The major predicted flux flows through the metabolic network under mixotrophic condition are shown in Figure 5. A small flux was observed through TCA cycle that suggests the ability of alga to undertake an energy efficient pathway for growth during mixotrophic growth.
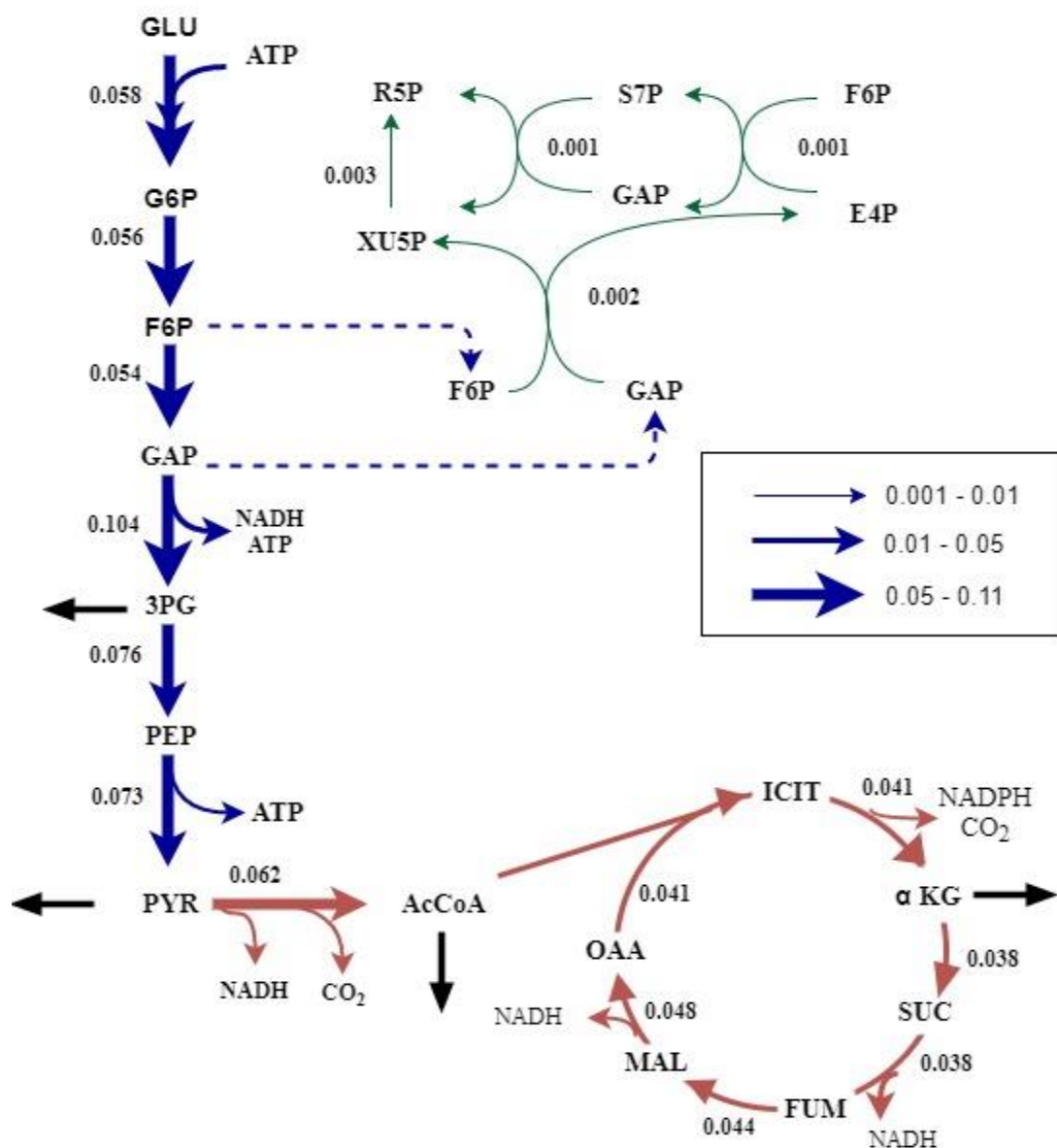
Figure 4 : Heterotrophic metabolic flux map

The flux values are shown near the arrows and also represented by the thickness of arrows. The reactions in green, blue and red colour represents oxidative pentose phosphate pathway, gluconeogenesis and TCA cycle respectively.
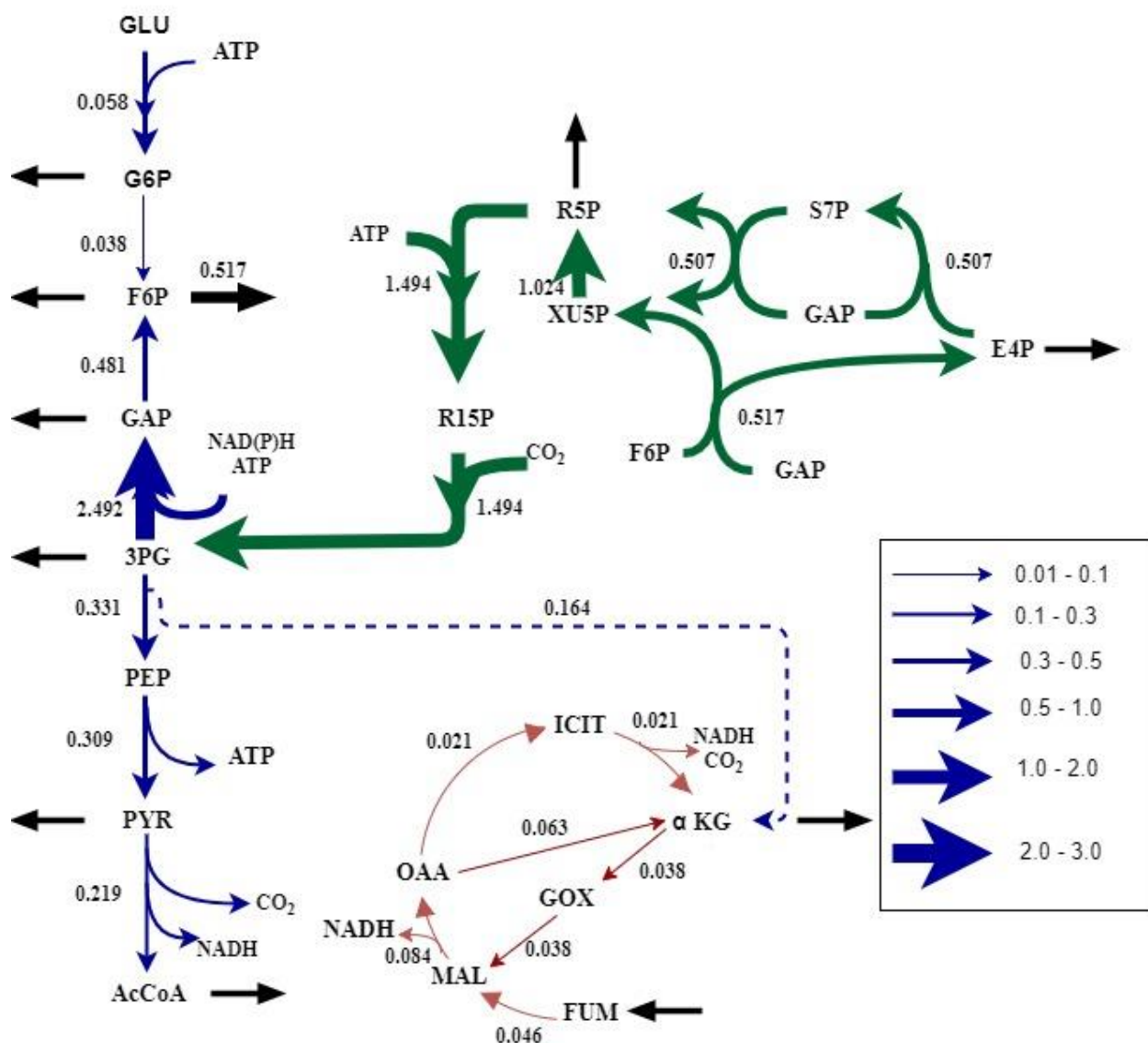
Figure 5 : Mixotrophic metabolic flux map

The flux values are shown near the arrows and also represented by the thickness of arrows. The reactions in green, blue and red colour represents reductive pentose phosphate pathway, gluconeogenesis and TCA cycle respectively.

5.5 Validation of the model

The metabolic model was validated by comparing the *in silico* predictions of flux through metabolic network and their ratios with experimental results reported in literature. Under

phototrophic condition, the growth rate predicted by flux balance analysis of the metabolic model seemed to be very close to the growth rate observed in experiment (given in section 6.2). The predicted photosynthetic quotient (moles of oxygen released per mole of carbon dioxide fixed) was found to be 1.12 and falls within the typical range found in algae, i.e., 1.0 to 1.8 (Burris, 1981). In addition, the value of CEF/(CEF+LEF) where CEF and LEF are the fluxes through cyclic electron flow and linear electron flow respectively, was found to be 0.042 by simulation of the metabolic model under phototrophic condition and the predicted value is in agreement with the experimentally reported value for *Nannochloropsis gaditana* by Simionato et al.(Simionato et al., 2013). In other words, the predicted value of ratio of PSII/PSI (i.e. the ratio of metabolic flux through photosystem II to that through photosystem I) was 0.95 which seemed to be close to the value experimentally reported for *N. gaditana* (Simionato et al., 2013). The lower growth rate under heterotrophic condition and higher growth rate under mixotrophic condition are in agreement with that reported for *Nannochloropsis* sp. in literature (Cheirsilp & Torpee, 2012; Das et al., 2011; Fang et al., 2004; Marudhupandi et al., 2016).

5.6 Conclusions

A genome-scale metabolic model for *M. gaditana CCMP526* was developed and simulated under different trophic conditions using flux balance analysis to predict the metabolic fluxes of intracellular metabolites. The predicted values of growth rate of the alga and other parameters seemed to be in agreement with experimental values reported in literature and thus, the metabolic model for the alga is validated.

Chapter 6

Physiological Studies of *M. gaditana CCMP526*

6.1 Introduction

The flux balance analysis (FBA) of the genome-scale metabolic model for *M. gaditana CCMP526* requires different physiological parameters to formulate biomass equation and to validate the metabolic model. The determination of metabolic flux in the metabolic reconstruction using FBA depends on the biomass formation equation which is derived from the biomass composition of the alga. FBA predicts the metabolic fluxes at steady state, i.e. total amount of metabolite being produced is equal to the total metabolite being consumed (Orth, Thiele, & Palsson, 2010). Therefore, we assume the algal system is at steady state while applying FBA, so we need to ensure that the alga is in exponential phase and under no stress while collecting sample for measurement. Hence studies were conducted to find the effect of different nitrate concentration and buffer (pH 8) on growth of the alga.

6.2 Growth study

M. gaditana CCMP526 was cultivated in f/2 medium with an initial cell concentration of $2*10^6$ cells mL^{-1}. The growth curve observed for the alga is shown in Figure 6. The growth kinetics parameters were calculated from the observed growth curve (calculations are shown below). The obtained experimental values of specific growth rate and corresponding doubling time are in agreement with the previous reports (Boussiba et al., 1985, 1987).
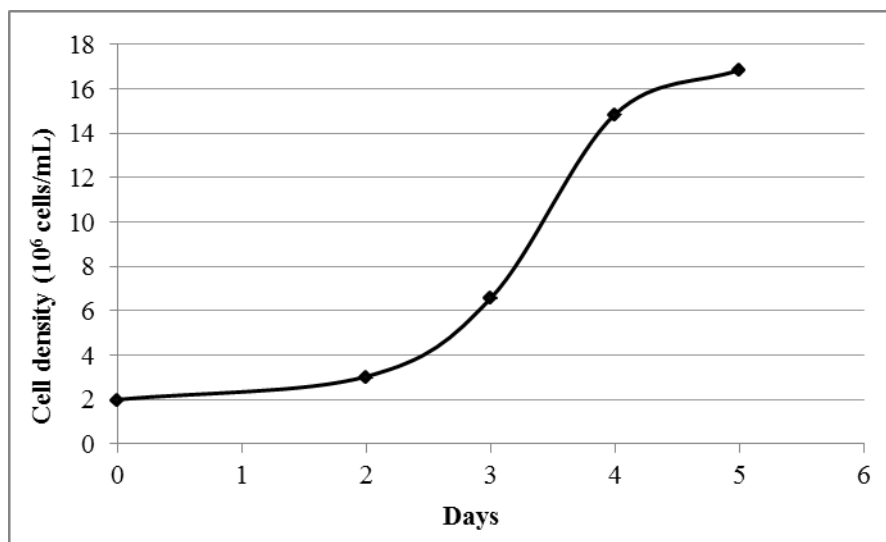
Figure 6: Growth of *M. gaditana CCMP526* in f/2 medium.

Data are means of 3 independent biological replicates ± standard error.

Growth kinetics:

Specific growth rate, $\mu = 0.814 \pm 0.02$ day^{-1} $= 0.033 \pm 0.001$ h^{-1}

Doubling time, $t_d = \ln2/\mu = 0.85 \pm 0.03$ days

Division per day $= 1/t_d = 1.17 \pm 0.04$

6.2.1 Selection of media for cultivation

Growth curves and parameters of *M. gaditana* cultivated in different media are given in Figure 7 and Table 8 respectively. Concentration of nitrate in f/2 medium was 0.88 mM. A higher concentration of nitrate, i.e. 17 mM nitrate, increased the cell density significantly during the exponential phase of cultivation. In addition, buffering the medium with Tris-Cl and thus maintaining a pH of 8 throughout the cultivation of culture prevented the culture turning pale throughout the exponential phase. A high cell density of $16.7*10^6$ cells/mL was reached earlier, i.e. on the 3rd day of cultivation, in 17 mM nitrate containing Tris-Cl buffer f/2 medium, as compared to that in all other media. So f/2 medium with 40 mM Tris-Cl buffer and 17mM sodium nitrate was selected for further experiments.
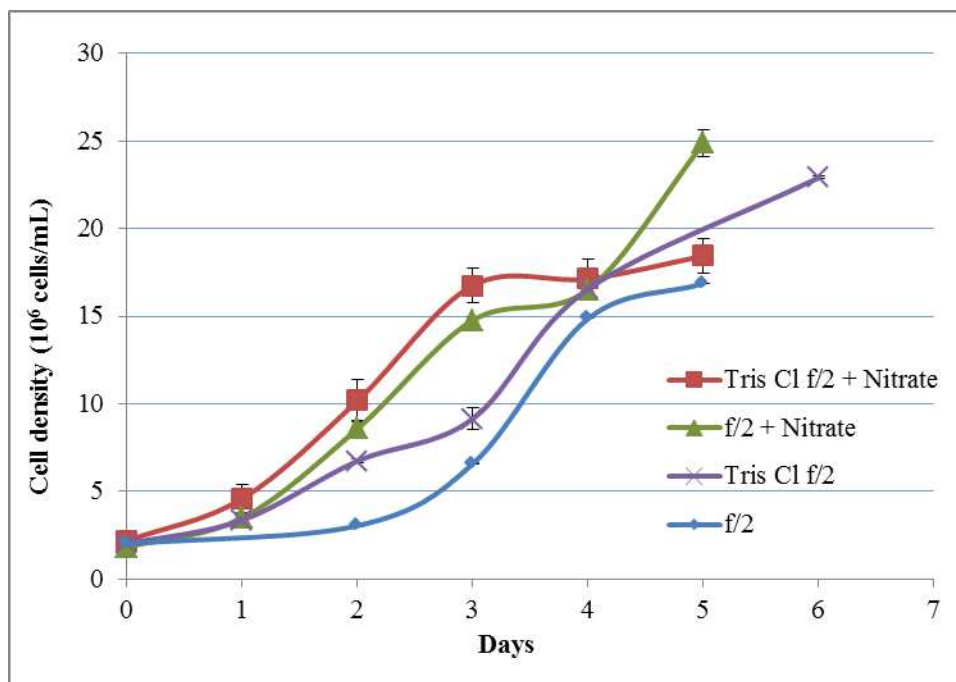
Figure 7: Growth curve of *M. gaditana* in media with buffer.

Data are means of 3 independent biological replicates ± standard error

.

Table 8 : Growth parameters of *M. gaditana* cultivated in different media.

Data are means of 3 independent biological replicates ± standard error

Media of sample	Specific growth rate (day-1)	Doubling time (hr)	Maximal biomass (10^6 cells/mL)
Tris Cl f/2 + 17 mM nitrate	0.83 ± 0.06	20.2 ± 1.68	16.75 ± 1.02
f/2 + 17 mM nitrate	0.92 ± 0.03	18.1 ± 0.54	14.77 ± 0.22
Tris Cl f/2	0.71 ± 0.04	23.5 ± 1.15	16.55 ± 0.30
f/2	0.83 ± 0.03	20.0 ± 0.72	14.86 ± 0.11

6.2.2 Effect of different nitrate concentrations on growth

Growth parameters of *M. gaditana* cultivated in buffered f/2 media with different nitrate concentration are given in Table 9. Specific growth rates of more or less same value were observed in f/2 buffered medium containing nitrate concentration less than 0.3mM. However, specific growth rate of the alga increase with the increase in nitrate concentration to 0.9mM, i.e. the concentration of nitrate in f/2 medium. However, the alga cultivated in media with 0.9 and

0.3 mM nitrate turned pale during the exponential phase and that in media with 0.1 mM and 0.05 mM nitrate turned pale in the beginning of exponential phase of its cycle. The alga that was cultivated in medium with 17mm nitrate seemed healthy without turning pale throughout the exponential phase. Therefore, f/2 medium with 17 mM nitrate was used for further experiments.

Table 9: Growth parameters of *M. gaditana* cultivated in media with different nitrate concentration.

Data are means of 3 independent biological replicates ± standard error

Nitrate conc. of sample	Specific growth rate (day-1)	Doubling time (hr)	Divisions per day
0.05 mM	0.45 ± 0.02	37.3 ± 1.3	0.64 ± 0.02
0.1 mM	0.50 ± 0.05	33.9 ± 4.0	0.72 ± 0.08
0.3 mM	0.50 ± 0.01	33.7 ± 0.7	0.71 ± 0.01
0.9 mM	0.71 ± 0.04	23.5 ± 1.15	1.03 ± 0.05
17 mM	0.83 ± 0.06	20.2 ± 1.7	1.20 ± 0.09

6.3 Sugar and protein estimation

The sugar content of *M. gaditana* during the exponential phase of growth was estimated using the phenol sulphuric acid method and the standard curve is given in Figure 12 (Appendix). The protein content of culture was also measured using the Bradford's reagent kit after cell lysis using sodium hydroxide treatment and the standard curve is given in Figure 13 (Appendix). The sugar and protein content of *M. gaditana* in its steady state are tabulated below. The obtained experimental values of sugar content and protein content are in their range reported for *Nannochloropsis* sp.(Sukenik, Carmeli, & Berner, 1989).

Table 10: Sugar and protein estimation of *M. gaditana*.

Data are means of 3 independent biological replicates ± standard error

Sugar content	Protein content
1.27 ± 0.02 pg/cell	2.23 ± 0.18 pg/cell

6.3.1 Sugar content profile

The sugar content of *M. gaditana CCMP526* was estimated using phenol sulphuric acid method and the average of the results of three experiments is shown in Figure 8. The sugar content seems to be in range when compared with that reported by Radakovits et al. (2012). It was observed that the sugar content of the culture increased during the log phase of the growth cycle and decreased towards the stationary phase. The highest sugar content was attained during the log phase.
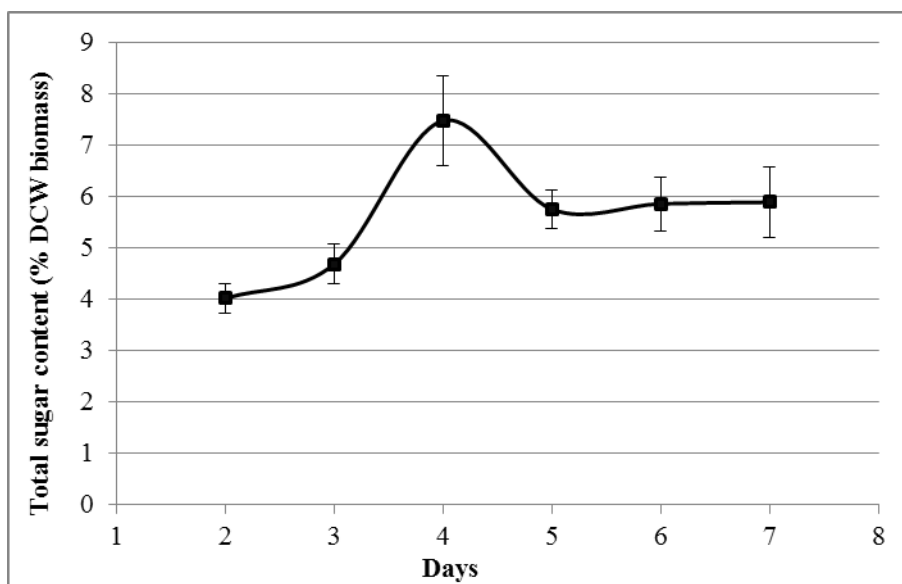
Figure 8 : Time profile of sugar content of *M. gaditana CCMP526.*

Data are means of 3 independent biological replicates ± standard error

6.4 Estimation of fatty acid profile

The fatty acid composition was qualitatively estimated for *M. gaditana CCMP526.* The 37 component Fatty Acid Methyl Esters (FAME) Mix was used as a standard for fatty acid measurement. The chromatogram obtained for 37 component FAME mix is given in Figure 14 and Table 29 (Appendix). The peaks were identified and quantified in the chromatogram obtained for *M. gaditana* using GC-MS method that is given in Figure 9. The main fatty acids present in the biomass of *M. gaditana* were C14:0, C16:0, C16:1, C18:0, C18:1n9c, C18:1n9t, C20:3n3, C24:1n9. The components were identified based on the spectrum of component obtained in MS and the retention time of the component. The approximate retention time taken

for most of the components was analysed using the standard. The fatty acid profile is required for calculating the biomass formation equation involved in the metabolic model. The qualitative analysis of fatty acids in *Microchloropsis gaditana* sp. seems to be in agreement with the reported fatty acid composition for *Nannochloropsis* sp. (Mourente et al., 1989).
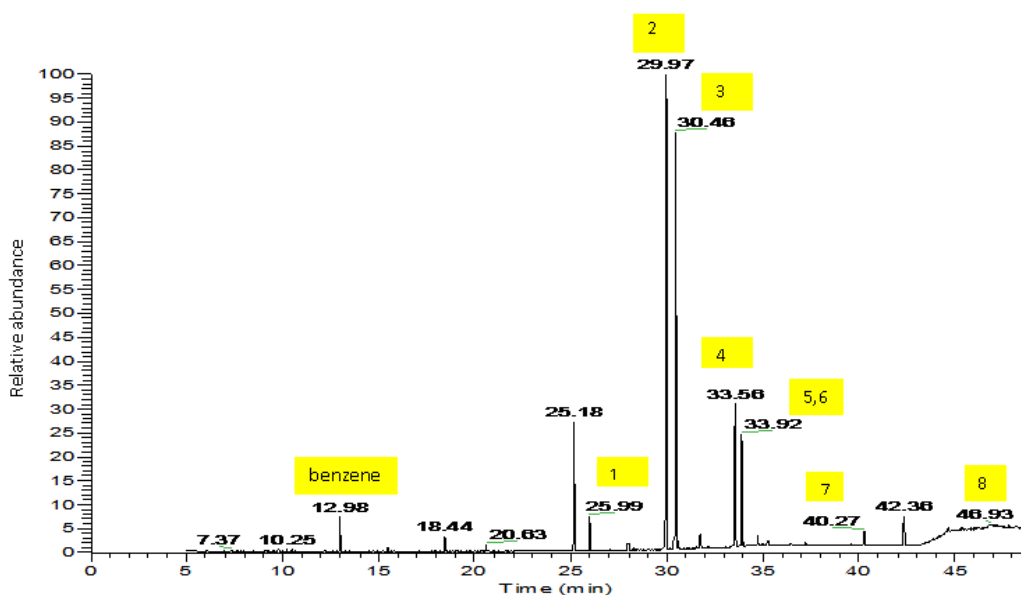
Figure 9 : Chromatogram of fatty acid profile of *M. gaditana*

Components labelled as follows 1(C14:0), 2(C16:0), 3(C16:1), 4(C18:0), 5 (C18:1n9c), 6 (C18:1n9t), 7(C20:3n3), 8(C24:1n9). Other peaks are not fatty acids.

6.5 Preliminary study for genetic transformation of *M. gaditana CCMP526*

6.5.1 Growth studies of *M. gaditana* in f/2 medium and TAP medium

The growth of the culture was monitored for *M. gaditana* in f/2 medium and TAP medium with different inoculum sizes of 1%, 5% and 10% of volume of medium, i.e., cell concentration of $9.98*10^5$ cells mL^{-1}, $2.082 * 10^6$ cells mL^{-1} and $3.034 * 10^6$ cells mL^{-1} respectively. The culture did not grow in TAP medium whereas it grew well in f/2 medium. The growth curves of *gaditana* with inocula of different size are given in Figure 7.
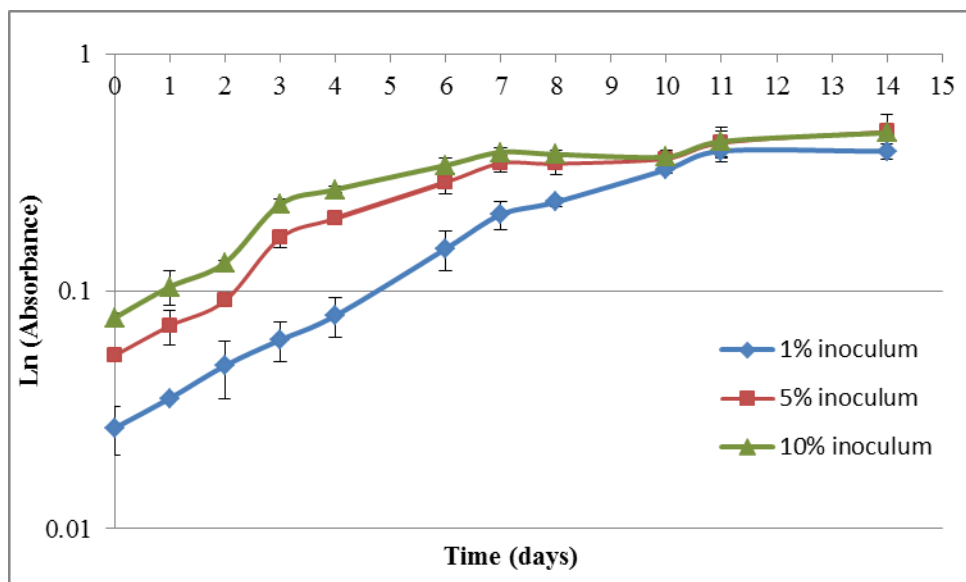
Figure 10 : Growth curve of *M. gaditana* in f/2 medium with different initial cell density.

Data are means of 3 independent biological replicates ± standard error

The preferred inoculum size for the genetic transformation is the one for which an early log phase is observed. However, the log phase seems to begin at the same time for all the cultures inoculated with different inoculum sizes. A cell number of around 10^7 cells mL^{-1} in the exponential phase is preferred for genetic transformation. The cell concentration reached the cell concentration of 10^7 cells mL^{-1} earlier during the growth of culture inoculated with an inoculum size of 10% of volume of medium, whereas the other cultures did not reach the cell concentration of 10^7 cells mL^{-1} during their exponential phase. However, there was no significant difference in the growth rate was observed for the cultures inoculated with different inoculums size. The statistics of growth kinetics are presented in Table 11. The cell concentrations (cell mL^{-1}) were calculated from optical density of culture inoculated with different inoculum sizes at a wavelength of 750nm (Lopes & Vasconcelos, 2011).The values of cell concentration of culture inoculated with different inoculums sizes are given in Table 3. So the inoculum size of 10% can be used for growing culture in f/2 medium to be used for genetic transformation.

Table 11: Growth kinetics of *M. gaditana* with different inoculum sizes.

Data are means of 3 independent biological replicates ± standard error

Growth kinetics	1% inoculum	5% inoculum	10% inoculum
Specific growth rate, μ (day^{-1})	0.31 ± 0.03	0.37 ± 0.04	0.36 ± 0.03
Doubling time, t_d (days)	2.32 ± 0.29	1.92 ± 0.18	1.99 ± 0.21
Divisions per day	0.44 ± 0.05	0.53 ± 0.05	0.51 ± 0.06
Maximum biomass in log phase (10^6 cells mL^{-1})	5.99 ± 1.15	6.65 ± 1.15	9.25 ± 1.15

6.5.2 Antibiotic sensitivity study

The antibiotic sensitivity studies of *M. gaditana* exposed to chloramphenicol, kanamycin and hygromycin were carried out at different concentrations (30μg/mL, 100μg/mL and 300μg/mL). Kanamycin is an antibiotic that binds to the 30S ribosomal unit whereas chloramphenicol affects 50S ribosomal subunit to restrict protein synthesis in the cell and thus they inhibit cell growth. Hygromycin restricts protein synthesis by disrupting translocation at the 70S ribosome. The results are shown in Figure 11. It was observed that the culture was resistant to the antibiotic kanamycin at all the concentrations whereas it was sensitive to chloramphenicol at all concentrations. The resistance of two microalgae that belong to *Nannochloropsis* sp. against kanamycin has been reported (Galloway, 1990) and we found it holds true for *Microchloropsis gaditana CCMP526*. The culture seems to be sensitive towards hygromycin at 300 μg/mL, though cells seemed to grow in media containing hygromycin at concentrations of 10 and 100 μg/mL from the 11th day onwards. Use of selectable marker conferring resistance to hygromycin in genetic transformation method for *Nannochloropsis* sp. (strain W2J3B) was reported (Kilian et al., 2011). Thus, the chloramphenicol resistance gene can also be used in the vector to be used for genetic transformation of *M. gaditana CCMP526* as the genetically transformed strains can be selected based on the expression of the resistance gene while growing in a chloramphenicol

containing medium. It was reported that marine algae of *Nannochloropsis* sp. are sensitive to antibiotics such as streptomycin and erythromycin (Galloway, 1990).
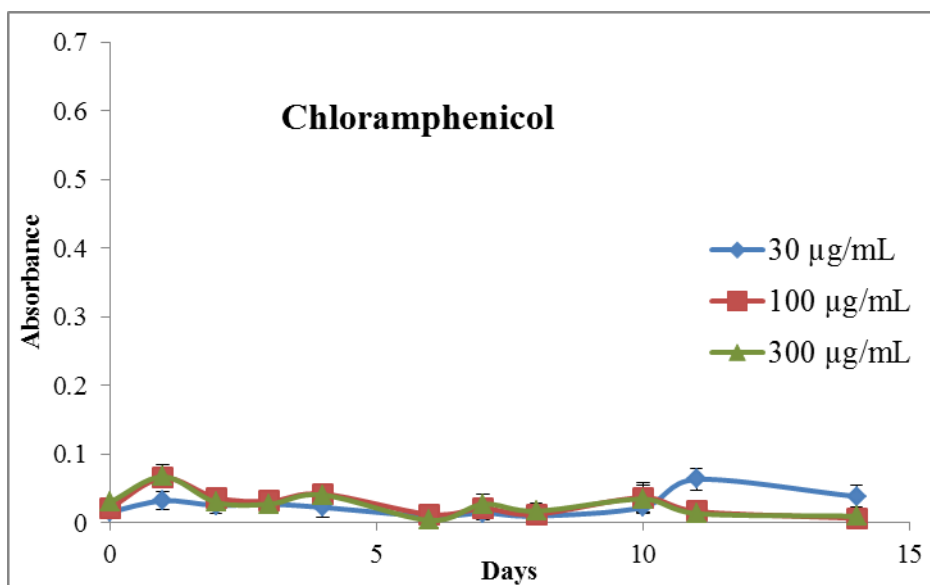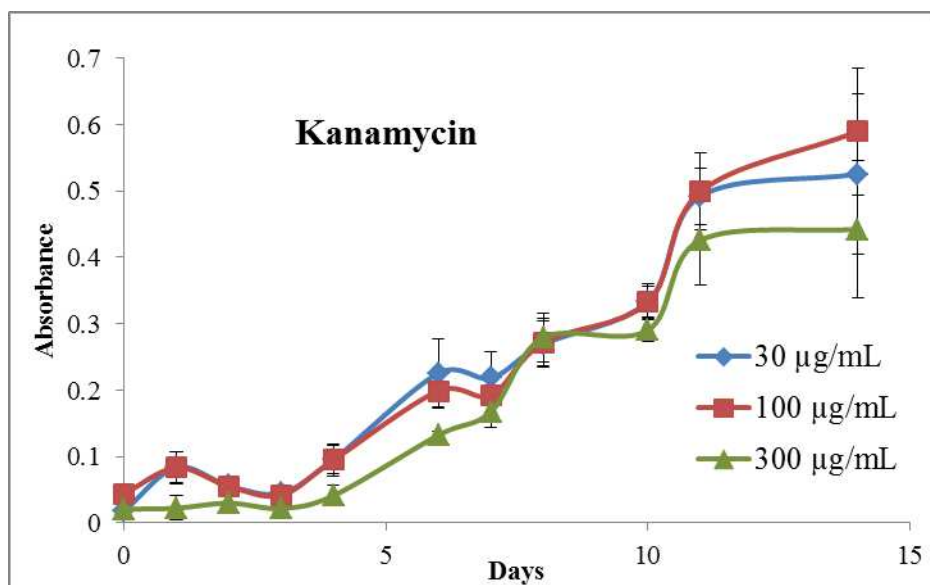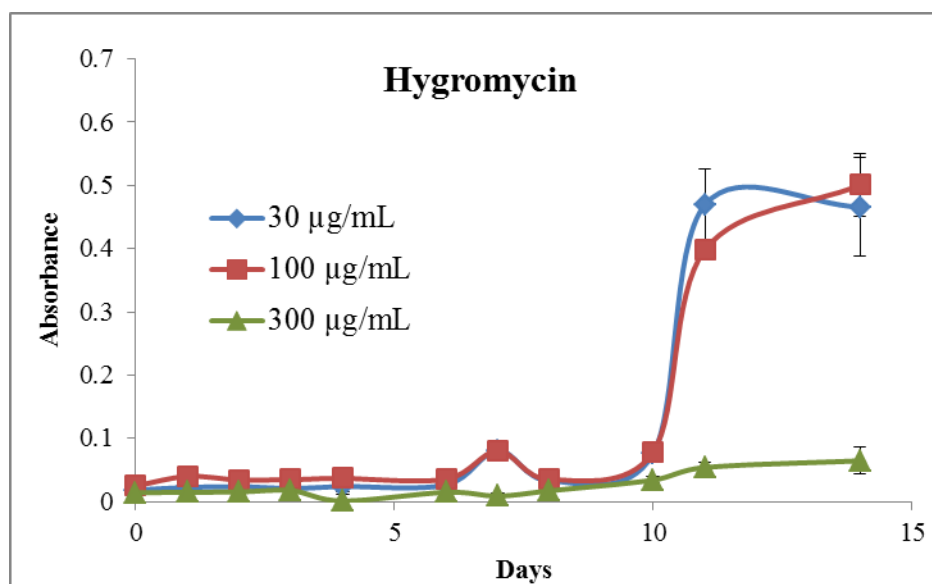
Figure 11 : Antibiotic sensitivity study on *M. gaditana*

Data are means of 3 independent biological replicates ± standard error

6.6 Conclusions

The growth rate of *M. gaditana CCMP526* was determined and contributed towards the validation of the metabolic model developed. The sugar content and the protein content of the alga under phototrophic condition during exponential phase was determined and used to formulate the biomass formation equation. The qualitative analysis of fatty acid profile of the alga ensures that there was no significant variation in biomass composition of the strain from that of *Microchloropsis salina,* which was used in the derivation of coefficients in biomass formation equation in metabolic model. Some preliminary studies on inoculum concentration and antibiotic sensitivity, which are required for genetic transformation of *M. gaditana,* were also carried out.

Chapter 7

Conclusions

A curated pathway database of *M. gaditana CCMP526,* MgdCyc, was developed using its genome sequence, genome annotation and experimental information related to the genus of *Microchloropsis* and *Nannochloropsis*. Since the genome annotation is very limited to an extent that even the reactions involved in the central carbon metabolism were not GPR (gene-protein-reaction) associated, development of flux consistent model has undergone an extensive curation. The developed biochemical pathway database features 141 metabolic pathways with associated genes, enzymes and metabolites with a user-friendly Pathway Tools graphical interface. The MgdCyc database can be downloaded as flat file from Biocyc distribution of pathway/genome database or can be accessed with Pathway Tools web. As a part of curation of the pathway database, genes of sixty one missing enzymes were identified in the genome sequence of the alga.

A genome-scale metabolic model was developed for *M. gaditana CCMP526* was developed by curating the metabolic network generated from the pathway database. Analysis of the metabolic model was carried out to validate the model and to understand the metabolic behaviour of the alga. Different trophic conditions such as photoautotrophic, heterotrophic and mixotrophic conditions were simulated to predict the fluxes of intracellular metabolites. The model was validated by comparing the predicted values of parameters such as photosynthetic coefficient, specific growth rate, ratio of cyclic and non-cyclic electron flow with that of experimental values. Some physiological studies of the alga were carried out to support the development of metabolic model and its validation.

7.1 Future Work

- The genome of *M. gaditana* can further undergo structural and functional annotation. The metabolic model developed guides to investigate more genes.

- Subcellular localization of enzymes in the pathway database can be undertaken as the genome sequencing project progresses

- ^{13}C-Metabolic flux analysis can be carried out to fill the gaps in the metabolic network and to validate the model (Zamboni, 2011).

- The metabolic model can be further analysed that bring insights into lipid accumulating ability of the microalga. Metabolic interventions can be predicted for optimal lipid synthesis using k-OptForce (Chowdhury, Zomorrodi, & Maranas, 2014).

APPENDICES

Table 12 : Details of genome annotation genome assembly v.1.2 of *M. gaditana* CCMP526 (Radakovits et al., 2012)

BioProject	PRJNA73791
Assembly	GCA_000240725.1
Level of genome assembly	Scaffold
Estimated genome size	33.987 Mb
Total number of genes	3557
Number of predicted proteins	3554
Number of hypothetical proteins	1582
Number of enzyme coding genes	391

Table 13 : Assigned reactions after curating probable enzyme matches

Probable enzyme match	Assigned reaction
2-succinyl-6-hydroxy- cyclohexadiene-1-carboxylic acid synthase 2-oxoglutarate decarboxylase	4.2.99.20 4.1.1.71;
3-hydroxyacyl-coa dehydrogenase	1.1.1.35 (21); 1.1.1.M19 (6); 4.2.1.74; 1.1.1.211(4)
3-ketoacyl- thiolase peroxisomal	2.3.1.16(5); 2.3.1.223;
Aaa family atpase	3.6.3.14
Aarf domain containing kinase 2	3.4.24.81; 2.7.1.154
Adenylyltransferase and sulfurtransferase	2.7.7.42
Aldehyde oxidase	1.2.3.1; 1.2.99.7;
Aldehyde reductase i	1.1.1.1(15); 1.1.1.2; 1.1.1.19; 1.1.1.21(5);
Alkyl sulfatase or beta-lactamase	1.14.11.M6; 3.1.6; 3.5.2.6
Amine oxidase	1.4.3.21; 1.4.3.22

Aminophospholipid-transporting p-type atpase	3.6.3.1
Beta- -endoglucanase	3.2.1.176; 3.2.1.4
Beta- -mannosyl-glycoprotein beta- -n-acetylglucosaminyltransferase	2.4.1.144
Cap-specific mrna (nucleoside-2 -o-)-methyltransferase 1	2.1.1.57
Choline dehydrogenase	1.1.99.1
Ctf2a like oxidoreductase	1.17.4.1 (4);
Ctf2a like oxidoreductase	1.17.4.1 (4);
Cyclin-dependent kinase 10	2.7.11.27; 2.7.11.20; 2.7.11.7;2.7.11.18; 2.7.11.1/2.7.11..12/2.7.11.22; 2.7.11.22
Dihydrolipoyllysine-residue acetyltransferase component 1 of pyruvate dehydrogenase complex	2.3.1.12
Diphthamide biosynthesis	6.3.1.14
Dolichol-phosphate mannosyltransferase subunit 3-like protein	2.4.1.109
E3 ubiquitin-protein ligase ubr4	2.3.2.23; 2.3.2.27
E3 ubiquitin-protein ligase ubr4	2.3.2.23; 2.3.2.28
E3 ubiquitin-protein ligase upl6	2.3.2.23; 2.3.2.29
E3 ubiquitin-protein ligase-like protein	2.3.2.23; 2.3.2.30
Erythromycin esterase	2.1.1.254
Ethanolamine kinase 1 isoform 1	2.7.1.82
Fatty acid desaturase	1.14.19.22; 1.14.19.1
Fatty acid elongase	6.2.1.3; 6.2.1.2;
Fe-Fe hydrogenase	1.12.7.2; 1.12.1.4;
Glutamine amidohydrolase-like protein	3.5.1.44
Glutamine amidotransferase	6.3.5.2
Glutathione peroxidase	1.11.1.9; 1.11.1.12
Glyceraldehyde-3-phosphate dehydrogenase	1.2.1.12; 1.2.1.59; 1.2.1.9; 1.2.1.13
Heavy metal p-type atpase	3.6.3.-
Hexose-6-phosphate dehydrogenase (glucose 1-dehydrogenase)	1.1.1.49; 1.1.1.388; 1.1.1.363;
Hnrnp arginine n-methyltransferase	2.1.1.319
Hydroxyacid oxidase	1.1.3.15

Hypoxanthine-guanine phosphoribosyltransferase-like protein	2.4.2.8
Indigoidine synthase a family protein	4.2.1.70
Inositol polyphosphate 5-phosphatase ocrl	3.1.3.56
Ketose-bisphosphate aldolase class-ii family protein	4.1.2.13
Lysine decarboxylase-like protein	4.1.1.18
Lysine ornithine decarboxylase	4.1.1.18; 4.1.1.17
Lysophosphatidylglycerol acyltransferase	2.3.1.-
Lysophospholipase-like 1	3.1.4.39, 3.1.1.5
Mg(2+) transport atpase protein c	3.6.3.2
N-terminal asparagine amidohydrolase	3.5.1.26; 3.5.1.38
Nad h dehydrogenase	1.6.5.3; 1.6.99.3; 1.6.5.9
Omega-6 fatty acid desaturase delta-12	1.14.19.6
P-type h+-atpase	3.6.3.6
Peptide methionine sulfoxide reductase	1.8.4.11; 1.8.4.12
Peptidyl-prolyl cis-trans isomerase d-like protein	5.2.1.8
Peptidyl-prolyl cis-trans isomerase fkbp2-like protein	5.2.1.8
Peptidyl-prolyl cis-trans isomerase-like protein	5.2.1.8
Phosphatidylinositol- -trisphosphate 5-phosphatase 1	3.1.3.86;
Phosphoglycerate bisphosphoglycerate mutase family protein	5.4.2.11
Phosphoglycerate mutase	5.4.2.12; 5.4.2.11
Poly rna polymerase	2.7.7.6; 2.7.7.48
Protein arginine n-methyltransferase	2.1.1.318;2.1.1.319; 2.1.1.320
Protein-s-isoprenylcysteine o-methyltransferase-like protein	2.1.1.100
Purple acid phosphatase isoform b2	3.1.3.2
Putative palmitoyltransferase zdhhc11-like protein	2.3.1.225
Putative serca-type calcium atpase	3.6.3.8
Putative tyrosinase-like protein in chromosome	1.2.4.1; 1.2.5.1
Receptor-interacting serine-threonine kinase 4	2.7.11.30

Ribulose- -bisphosphate carboxylase oxygenase small subunit n-methyltransferase i	2.1.1.127
rRNA (guanine-n -)-methyltransferase	2.1.1.171
Sphingolipid delta-4 desaturase	1.14.19.17
Sphingosine-1-phosphate lyase	4.1.2.27
Threonine aldolase	4.1.2.48
Trehalose synthase	2.4.1.245; 5.4.99.16;
tRNA pseudouridine synthase	5.4.99.25; 5.4.99.28; 5.4.99.12; 5.4.99.27; 5.4.99.26;
Ubiquitin carboxyl-terminal hydrolase 10	3.4.19.12
Ubiquitin carboxyl-terminal hydrolase 24	3.4.19.12
Ubiquitin ligase e3	6.2.1.45; 2.3.2.26

Table 14 : Pathways that are reported to be present in *Microchloropsis* sp.

Pathways	Reference
Acetyl-CoA biosynthesis II	(Li et al., 2014)
Biosynthesis of amino acids	(Radakovits et al., 2012)
Calvin cycle	(Li et al., 2014; Radakovits et al., 2012)
Carotenoid Biosynthesis	(Radakovits et al., 2012)
Fatty acid biosynthesis	(Radakovits et al., 2012)
Glycolysis	(Li et al., 2014)
Glyoxylate cycle	(Vieler et al., 2012)
Lipid biosynthetic process	(Radakovits et al., 2012)
Nitrogen compound metabolic process	(Radakovits et al., 2012)
Pentose phosphate pathway	(Alboresi et al., 2016)
Photosynthesis	(Radakovits et al., 2012)
Sterol Synthesis	(Radakovits et al., 2012)
TCA cycle	(Li et al., 2014)

Tetrapyrrole Synthesis	(Radakovits et al., 2012)
β-oxidation	(Li et al., 2014)

Table 15: Filled pathway holes in *M. gaditana CCMP526*

The pathway holes were filled using Pathway Hole Filler where the database, YeastCyc was used to train the Bayes classifier.

Hole-filler	Hole EC#	P(has-function)	All functions of hole-filler	Pathway(s) requiring this reaction
ARGC	2.7.2.8	0.977	N-acetyl-gamma-glutamyl-phosphate/N-acetyl- gamma-aminoadipyl-phosphate reductase	L-arginine biosynthesis III (via N-acetyl-L-citrulline), L-arginine biosynthesis II (acetyl cycle), L-ornithine biosynthesis I
ARGC	1.2.1.-	0.977	N-acetyl-gamma-glutamyl-phosphate/N-acetyl- gamma-aminoadipyl-phosphate reductase	L-lysine biosynthesis V
UBE2N	none	0.989	ubiquitin-conjugating enzyme E2 N	protein ubiquitylation
UBE2N	none	0.978	ubiquitin-conjugating enzyme E2 N	protein ubiquitylation
MTNA	4.2.1.109	0.992	methylthioribose-1-phosphate isomerase	S-methyl-5-thio-alpha-D-ribose 1-phosphate degradation
MTNA	2.7.1.100	0.974	methylthioribose-1-phosphate isomerase	S-methyl-5'-thioadenosine degradation I
NGA_0114901	1.1.1.178	0.933	short chain dehydrogenase	L-isoleucine degradation I
NGA_0126900	3.1.1.23	0.933	esterase lipase thioesterase family	triacylglycerol degradation

			protein	
DES	1.14.18.4	0.979	omega-6 fatty acid desaturase delta-12	ricinoleate biosynthesis
DES	1.14.18.4	0.986	omega-6 fatty acid desaturase delta-12	ricinoleate biosynthesis, hydroxylated fatty acid biosynthesis (plants)
DES	1.14.18.4	0.979	omega-6 fatty acid desaturase delta-12	hydroxylated fatty acid biosynthesis (plants)
DES	1.14.19.25	0.925	omega-6 fatty acid desaturase delta-12	hydroxylated fatty acid biosynthesis (plants)
DES	1.14.19.34	0.979	omega-6 fatty acid desaturase delta-12	dimorphecolate biosynthesis
DES	1.14.19.35	0.963	omega-6 fatty acid desaturase delta-12	(7Z,10Z,13Z)-hexadecatrienoate biosynthesis
DES	1.14.19.25	0.925	omega-6 fatty acid desaturase delta-12	alpha-linolenate biosynthesis I (plants and red algae)
DES	1.14.19.35	0.963	omega-6 fatty acid desaturase delta-12	alpha-linolenate biosynthesis I (plants and red algae)
MET3	2.7.1.25	0.998	sulfate adenylyltransferase	sulfate activation for sulfonation
PDC	2.2.1.6	0.935	pyruvate decarboxylase	L-isoleucine biosynthesis IV, L-isoleucine biosynthesis III, L-isoleucine biosynthesis II, L-isoleucine biosynthesis I (from threonine)
NGA_0209900	1.14.19.4	0.997	delta 5 fatty acid desaturase	arachidonate biosynthesis IV (8-detaturase, lower eukaryotes)

NGA_020 9900	1.14.19.47	0.998	delta 5 fatty acid desaturase	arachidonate biosynthesis I (6-desaturase, lower eukaryotes), dicranin biosynthesis
NGA_020 9900	1.14.19.3	0.912	delta 5 fatty acid desaturase	(4Z,7Z,10Z,13Z,16Z)-docosa-4,7,10,13,16-pentaenoate biosynthesis (6-desaturase)
NGA_020 9900	1.14.19.3	0.912	delta 5 fatty acid desaturase	gamma-linolenate biosynthesis II (animals)
GSR	1.2.1.-	0.977	glutathione reductase (NADPH)	L-isoleucine biosynthesis V, L-isoleucine degradation I
NGA_037 3902	4.1.3.1	0.903	malate synthase	TCA cycle V (2-oxoglutarate:ferredoxin oxidoreductase), TCA cycle IV (2-oxoglutarate decarboxylase), glyoxylate cycle
NGA_043 4501	none	0.948	amine oxidase	10,13-epoxy-11-methyl-octadecadienoate biosynthesis
LDHA	1.1.1.29	0.91	D-lactate dehydrogenase	photorespiration
CCBL	2.6.1.79	0.956	kynurenine-oxoglutarate transaminase / cysteine-S-conjugate beta-lyase	L-tyrosine biosynthesis III, L-tyrosine biosynthesis II, L-phenylalanine biosynthesis II
CCBL	2.6.1.17	0.925	kynurenine-oxoglutarate transaminase / cysteine-S-conjugate beta-lyase	L-lysine biosynthesis I
CCBL	2.6.1.-	0.979	kynurenine-oxoglutarate transaminase / cysteine-S-conjugate beta-lyase	S-methyl-5-thio-alpha-D-ribose 1-phosphate degradation

Table 16: Filled pathway holes in *M. gaditana CCMP526*

The pathway holes were filled using Pathway Hole Filler where the database, EcoCyc was used to train the Bayes classifier.

Hole-filler	Hole EC#	P(has-function)	All functions of hole-filler	Pathway(s) requiring this reaction
UBE2N	2.3.2.24	0.922	ubiquitin-conjugating enzyme E2 N	protein ubiquitylation
PTER	1.3.1.9	0.982	peroxisomal trans-2-enoyl-CoA reductase	palmitoleate biosynthesis I (from (5Z)-dodec-5-enoate)
PTER	1.3.1.9	0.982	peroxisomal trans-2-enoyl-CoA reductase	superpathway of mycolate biosynthesis, mycolate biosynthesis
PTER	1.3.1.34	0.955	peroxisomal trans-2-enoyl-CoA reductase	fatty acid beta-oxidation V (unsaturated, odd number, di-isomerase-dependent)
PTER	1.3.1.9	0.982	peroxisomal trans-2-enoyl-CoA reductase	cis-vaccenate biosynthesis
PTER	1.3.1.9	0.982	peroxisomal trans-2-enoyl-CoA reductase	(5Z)-dodec-5-enoate biosynthesis
NGA_064 0610	2.3.1.199	0.998	fatty-acyl	very long chain fatty acid biosynthesis I
NGA_064 0610	2.3.1.199	0.998	fatty-acyl	stearate biosynthesis I (animals and fungi)
NGA_064 0610	2.3.1.199	0.998	fatty-acyl	juniperonate biosynthesis
NGA_064 0610	2.3.1.199	0.998	fatty-acyl	hydroxylated fatty acid biosynthesis (plants)
NGA_064 0610	2.3.1.199	0.998	fatty-acyl	hydroxylated fatty acid biosynthesis (plants)
DES	1.14.19.23	1	omega-6 fatty acid desaturase delta-12	(7Z,10Z,13Z)-hexadecatrienoate biosynthesis
PDC	2.2.1.6	0.979	pyruvate decarboxylase	L-valine biosynthesis

RFBB	5.1.3.2	0.996	dTDP-glucose 4,6-dehydratase	D-galactose degradation V (Leloir pathway)
NGA_020 9900	1.14.19.44	0.955	delta 5 fatty acid desaturase	arachidonate biosynthesis V (8-detaturase, mammals), arachidonate biosynthesis III (6-desaturase, mammals)
NGA_042 0002	none	0.955	glycine cleavage system h protein	photorespiration
GSR	none	0.932	glutathione reductase (NADPH)	TCA cycle VII (acetate-producers), TCA cycle II (plants and fungi), TCA cycle I (prokaryotic)
GSR	1.8.1.4	0.932	glutathione reductase (NADPH)	pyruvate decarboxylation to acetyl CoA
NGA_043 4501	2.1.1.-	1	amine oxidase	10,13-epoxy-11-methyl-octadecadienoate biosynthesis

Table 17 : Pathways added to the database, MgdCyc

Pathway ID in MgdCyc	Pathway name
HOMOCYSDEGR-PWY	L-cysteine biosynthesis III (from L-homocysteine)
PWY-922	mevalonate pathway I
PWY-7117	C4 photosynthetic carbon assimilation cycle, PEPCK type
PWY-7115	C4 photosynthetic carbon assimilation cycle, NAD-ME type
P21-PWY	pentose phosphate pathway (partial)
HISTSYN-PWY	L-histidine biosynthesis
ALANINE-VALINESYN-PWY	L-alanine biosynthesis I
PWY-4981	L-proline biosynthesis II (from arginine)
PWY-7560	methylerythritol phosphate pathway II
NONMEVIPP-PWY	methylerythritol phosphate pathway I
PWY-401	galactolipid biosynthesis I

PWY-4381	fatty acid biosynthesis initiation I
P105-PWY	TCA cycle IV (2-oxoglutarate decarboxylase)
GLUTORN-PWY	L-ornithine biosynthesis I
TRPSYN-PWY	L-tryptophan biosynthesis
PWY-5097	L-lysine biosynthesis VI
OXIDATIVEPENT-PWY	pentose phosphate pathway (oxidative branch)
NONOXIPENT-PWY	pentose phosphate pathway (non-oxidative branch)
ILEUSYN-PWY	L-isoleucine biosynthesis I (from threonine)
PWY-101	photosynthesis light reactions
SAM-PWY	S-adenosyl-L-methionine biosynthesis
PHESYN	L-phenylalanine biosynthesis I
PWY-7400	L-arginine biosynthesis IV (archaebacteria)
HOMOSER-METSYN-PWY	L-methionine biosynthesis I
PWY-5971	palmitate biosynthesis II (bacteria and plants)
VALSYN-PWY	L-valine biosynthesis
LEUSYN-PWY	L-leucine biosynthesis
PWY-7388	octanoyl-[acyl-carrier protein] biosynthesis (mitochondria, yeast)
GLYCOLYSIS	glycolysis I (from glucose 6-phosphate)
PWY-6837	fatty acid beta-oxidation V (unsaturated, odd number, di-isomerase-dependent)
FASYN-ELONG-PWY	fatty acid elongation -- saturated
PWY66-391	fatty acid β-oxidation VI (peroxisome)
PWY-5989	stearate biosynthesis II (bacteria and plants)
PWY-4361	*S*-methyl-5-thio-α-D-ribose 1-phosphate degradation
PWY-5484	glycolysis II (from fructose 6-phosphate)
PWY-5136	fatty acid β-oxidation II (peroxisome)

PWY-6134	L-tyrosine biosynthesis IV
PWY-6368	3-phosphoinositide degradation
PWY-381	nitrate reduction II (assimilatory)
PWY-6429	ricinoleate biosynthesis
PWY-7663	gondoate biosynthesis (anaerobic)
PWY0-862	(5Z)-dodec-5-enoate biosynthesis
PWY-7589	palmitoleate biosynthesis III (cyanobacteria)
HOMOSER-THRESYN-PWY	L-threonine biosynthesis
PWY-5344	L-homocysteine biosynthesis
PWY-6352	3-phosphoinositide biosynthesis
PWY-1042	glycolysis IV (plant cytosol)
PWY-5973	*cis*-vaccenate biosynthesis
PWY-5675	nitrate reduction V (assimilatory)
PWY-6599	guanine and guanosine salvage II
PWY490-4	L-asparagine biosynthesis III (tRNA-dependent)
CALVIN-PWY	Calvin-Benson-Bassham cycle
PWY-7725	arachidonate biosynthesis V (8-detaturase, mammals)
PWY-7344	UDP-D-galactose biosynthesis
PWY-181	photorespiration
PWY-7094	fatty acid salvage
PWY-7590	(7Z,10Z,13Z)-hexadecatrienoate biosynthesis
PWY-7587	oleate biosynthesis III (cyanobacteria)
PWY-7036	very long chain fatty acid biosynthesis II
GLYOXYLATE-BYPASS	glyoxylate cycle
PWY-6282	palmitoleate biosynthesis I (from (5Z)-dodec-5-enoate)

PWY3DJ-12	ceramide *de novo* biosynthesis
PWY-6754	*S*-methyl-5'-thioadenosine degradation I
PWY66-21	ethanol degradation II
PWY-7726	(4Z,7Z,10Z,13Z,16Z)-docosa-4,7,10,13,16-pentaenoate biosynthesis (6-desaturase)
PWY-6970	acetyl-CoA biosynthesis II (NADP-dependent pyruvate dehydrogenase)
PWY-7691	10,13-epoxy-11-methyl-octadecadienoate biosynthesis
PWY-6000	γ-linolenate biosynthesis II (animals)
PWY-5997	α-linolenate biosynthesis I (plants and red algae)
PWY-6001	linoleate biosynthesis II (animals)
PWY-46	putrescine biosynthesis III
PWY-6019	pseudouridine degradation
PWY-5080	very long chain fatty acid biosynthesis I
PWY-5340	sulfate activation for sulfonation
PWY-4081	glutathione-peroxide redox reactions
PYRUVDEHYD-PWY	pyruvate decarboxylation to acetyl CoA

Table 18 : Filled pathway holes in MgdCyc

The pathway holes were filled using Pathway Hole Filler where the databases, YeastCyc and EcoCyc were used to train the Bayes classifier.

Hole-filler	Hole EC#	P(has-function)	All functions of hole-filler	Pathway(s) requiring this reaction
RFBB	4.2.1.76	0.98	dTDP-glucose 4,6-dehydratase	UDP-L-rhamnose biosynthesis
NGA_0209900	1.14.19.44	0.96	delta 5 fatty acid desaturase	icosapentaenoate biosynthesis III (8-desaturase, mammals), icosapentaenoate

				biosynthesis II (6-desaturase, mammals)
NGA_0209900	1.14.19.4	1.00	delta 5 fatty acid desaturase	icosapentaenoate biosynthesis V (8-desaturase, lower eukaryotes)
NGA_0209900	1.14.19.3	0.91	delta 5 fatty acid desaturase	icosapentaenoate biosynthesis II (6-desaturase, mammals)
NGA_0209900	1.14.19.47	0.99	delta 5 fatty acid desaturase	icosapentaenoate biosynthesis I (lower eukaryotes)
NGA_0429100	1.1.1.59	0.93	choline dehydrogenase	beta-alanine biosynthesis II
PGM	5.4.2.8	0.91	phosphoglucomutase	GDP-mannose biosynthesis
GLYA	none	0.99	glycine hydroxymethyltransferase	folate transformations II
PTER	1.1.1.10	0.933	peroxisomal trans-2-enoyl-CoA reductase	-
LDHA	1.1.1.81	0.955	D-lactate dehydrogenase	-

Table 19: BLAST results for enzymes involved in Glycolysis

Enzyme	Best hit	% identity	s. start	s. end	E value
Glucose-6-phosphate isomerase	NW_005803947.1	100.0	74858	72766	0
Enolase	NW_005803939.1	99.9	40911	43959	0
	NW_005803793.1	99.2	33529	31945	0

Table 20 : BLAST results for enzymes involved in TCA cycle

Enzyme name	Best hit	% identity	s. start	s. end	E value
Citrate synthase	NW_005803933.1	99.7	28739	26120	0
	NW_005803853.1	99.9	44371	42449	0
Isocitrate dehydrogenase	NW_005803939.1	100.0	79174	82265	0
2-oxoglutarate dehydrogenase E1 component	NW_005803894.1	99.8	5121	4042	0
	NW_005803767.1	99.9	20771	17077	0
Succinate dehydrogenase iron sulfur protein	NW_005803731.1	99.7	21098	19216	0
Succinate dehydrogenase flavoprotein subunit	NW_005803892.1	99.9	44959	41582	0
Succinate dehydrogenase subunit 4	NW_005803712.1	100.0	4366	3811	0
Succinate dehydrogenase cytochrome b subunit	NW_005803425.1	99.0	13685	12166	0
Fumarate hydratase	NW_005803860.1	100.0	64654	63189	0
	NW_005803848.1	99.5	23950	25777	0

Table 21 : BLAST results for enzymes involved in calvin cycle

Enzyme	Best hit	% identity	s. start	s. end	E value
Fructose- -bisphosphatase	NW_005803949.1	99.9	20017	21779	0
	NW_005802355.1	100.0	783	309	0

	NW_005803948.1	99.7	144885	143256	0
	NW_005802797.1	99.7	33	1039	0
Phosphoribulokinase	NW_005803849.1	98.9	53616	55206	0
	NW_005803390.1	99.1	5212	2576	0

Table 22: BLAST results for enzymes involved in Pentose phosphate pathway

Enzyme	Best hit	% identity	s. start	s. end	E value
6-phosphogluconolactonase	NW_005803829.1	100.0	15361	14271	0
Transketolase	NW_005803884.1	99.8	61958	58637	0
	NW_005803832.1	99.5	4314	7215	0
	NW_005803832.1	100.0	1	56	5.21E-21
	NW_005803471.1	99.7	1	1542	0
	NW_005803739.1	99.4	6541	4370	0
	NW_005803735.1	99.4	6250	4079	0
Transaldolase	NW_005803726.1	93.8	6835	7825	0
	NW_005803726.1	99.2	7085	7836	0
	NW_005803726.1	100.0	13566	13085	0
	NW_005803726.1	99.1	6892	7002	2.80E-50

NW_005803726.1	99.1	13359	13249	2.80E-50
NW_005803726.1	93.7	13163	13085	8.06E-26
NW_005803726.1	84.5	13500	13398	8.12E-21
NW_005803517.1	98.9	5187	3365	0
NW_005803920.1	100.0	24474	25701	0

Table 23 : Pathways in the database, MgdCyc with their pathway prediction score

Pathway prediction score is a number that indicates the strength of the evidence supporting the inference of the pathway and ranges from 0-1.0 where 1.0 means very strong evidence.

COMMON-NAME	FRAME	SCORE
(4Z,7Z,10Z,13Z,16Z)-docosa-4,7,10,13,16-pentaenoate biosynthesis (6-desaturase)	PWY-7726	0.381
(5Z)-dodec-5-enoate biosynthesis	PWY0-862	0.200
(7Z,10Z,13Z)-hexadecatrienoate biosynthesis	PWY-7590	0.267
cis vaccenate biosynthesis	PWY-5973	0.210
myo-inositol biosynthesis	PWY-2301	0.715
S-methyl-5-thio-α-D-ribose 1-phosphate degradation	PWY-4361	0.165
S-methyl-5'-thioadenosine degradation I	PWY-6754	0.325
10,13-epoxy-11-methyl-octadecadienoate biosynthesis	PWY-7691	0.400
2-oxoisovalerate decarboxylation to isobutanoyl-CoA	PWY-5046	0.560
3-phosphoinositide biosynthesis	PWY-6352	0.203
3-phosphoinositide degradation	PWY-6368	0.196
4-hydroxyphenylpyruvate biosynthesis	PWY-5886	0.236

acetyl-CoA biosynthesis II (NADP-dependent pyruvate dehydrogenase)	PWY-6970	0.400
acyl carrier protein activation	PWY-6012-1	0.423
adenine and adenosine salvage VI	PWY-6619	0.880
adenosine deoxyribonucleotides *de novo* biosynthesis	PWY-7227	0.373
adenosine nucleotides degradation II	SALVADEHYPOX-PWY	0.343
adenosine ribonucleotides *de novo* biosynthesis	PWY-7219	0.440
allantoin degradation to ureidoglycolate I (urea producing)	PWY-5697	0.440
alpha;-linolenate biosynthesis I (plants and red algae)	PWY-5997	0.400
arachidonate biosynthesis V (8-detaturase, mammals)	PWY-7725	0.238
C4 photosynthetic carbon assimilation cycle, NAD-ME type	PWY-7115	0.033
C4 photosynthetic carbon assimilation cycle, PEPCK type	PWY-7117	0.030
Calvin-Benson-Bassham cycle	CALVIN-PWY	0.317
CDP-diacylglycerol biosynthesis I	PWY-5667	0.212
ceramide *de novo* biosynthesis	PWY3DJ-12	0.300
chorismate biosynthesis from 3-dehydroquinate	PWY-6163	0.067
CMP phosphorylation	PWY-7205	0.440
D-myo-inositol (1,4,5)-trisphosphate degradation	PWY-6363	0.587
D-galactose degradation V (Leloir pathway)	PWY66-422	0.290
diacylglycerol and triacylglycerol biosynthesis	TRIGLSYN-PWY	0.247
docosahexaenoate biosynthesis III (6-desaturase, mammals)	PWY-7606	0.251
ethanol degradation II	PWY66-21	0.345
fatty acid β-oxidation II (peroxisome)	PWY-5136	0.186
fatty acid β-oxidation VI (peroxisome)	PWY66-391	0.160
fatty acid beta-oxidation V (unsaturated, odd number, di-isomerase-dependent)	PWY-6837	0.160

fatty acid biosynthesis initiation I	PWY-4381	0.081
fatty acid elongation -- saturated	FASYN-ELONG-PWY	0.160
fatty acid salvage	PWY-7094	0.267
fructose 2,6-bisphosphate biosynthesis	PWY66-423	0.880
galactolipid biosynthesis I	PWY-401	0.080
gamma;-linolenate biosynthesis II (animals)	PWY-6000	0.400
GDP-L-fucose biosynthesis II (from L-fucose)	PWY-6	0.440
glutaminyl-tRNA^{gln} biosynthesis via transamidation	PWY-5921	0.440
glutathione biosynthesis	GLUTATHIONESYN-PWY	0.440
glutathione-glutaredoxin redox reactions	GLUT-REDOX-PWY	0.423
glutathione-peroxide redox reactions	PWY-4081	0.728
glycerol-3-phosphate shuttle	PWY-6118	0.440
glycine biosynthesis I	GLYSYN-PWY	0.233
glycolysis I (from glucose 6-phosphate)	GLYCOLYSIS	0.219
glycolysis II (from fructose 6-phosphate)	PWY-5484	0.218
glycolysis III (from glucose)	ANAGLYCOLYSIS-PWY	0.236
glycolysis IV (plant cytosol)	PWY-1042	0.216
glyoxylate cycle	GLYOXYLATE-BYPASS	0.315
gondoate biosynthesis (anaerobic)	PWY-7663	0.200
guanine and guanosine salvage II	PWY-6599	0.212
guanosine deoxyribonucleotides *de novo* biosynthesis I	PWY-7226	0.373
heme biosynthesis I (aerobic)	HEME-BIOSYNTHESIS-II	0.267
heme degradation	PWY-5874	0.440
L-alanine biosynthesis I	ALANINE-VALINESYN-PWY	0.043
L-arginine biosynthesis IV (archaebacteria)	PWY-7400	0.129

L-arginine degradation X (arginine monooxygenase pathway)	ARGDEG-V-PWY	0.293
L-ascorbate degradation V	PWY-6959	0.000
L-asparagine biosynthesis III (tRNA-dependent)	PWY490-4	0.217
L-aspartate biosynthesis	ASPARTATESYN-PWY	0.239
L-aspartate degradation I	ASPARTATE-DEG1-PWY	0.239
L-citrulline degradation	CITRULLINE-DEG-PWY	0.401
L-cysteine biosynthesis III (from L-homocysteine)	HOMOCYSDEGR-PWY	0.000
L-glutamate degradation I	GLUTAMATE-DEG1-PWY	0.263
L-glutamate degradation IX (via 4-aminobutanoate)	PWY0-1305	
L-glutamine degradation I	GLUTAMINDEG-PWY	0.251
L-histidine biosynthesis	HISTSYN-PWY	0.040
L-homocysteine biosynthesis	PWY-5344	0.200
linoleate biosynthesis II (animals)	PWY-6001	0.421
lipoate biosynthesis and incorporation I	PWY0-501	0.440
L-isoleucine biosynthesis I (from threonine)	ILEUSYN-PWY	0.105
L-isoleucine degradation I	ILEUDEG-PWY	0.323
L-leucine biosynthesis	LEUSYN-PWY	0.144
L-lysine biosynthesis VI	PWY-5097	0.093
L-methionine biosynthesis I	HOMOSER-METSYN-PWY	0.130
long-chain fatty acid activation	PWY-5143	0.283
L-ornithine biosynthesis I	GLUTORN-PWY	0.088
L-phenylalanine biosynthesis I	PHESYN	0.126
L-proline biosynthesis II (from arginine)	PWY-4981	0.050
L-serine biosynthesis	SERSYN-PWY	0.293
L-threonine biosynthesis	HOMOSER-THRESYN-PWY	0.200

L-tryptophan biosynthesis	TRPSYN-PWY	0.088
L-tyrosine biosynthesis IV	PWY-6134	0.192
L-valine biosynthesis	VALSYN-PWY	0.141
L-valine degradation I	VALDEG-PWY	0.475
melibiose degradation	PWY0-1301	0.880
methylerythritol phosphate pathway I	NONMEVIPP-PWY	0.056
methylerythritol phosphate pathway II	PWY-7560	0.056
mevalonate pathway I	PWY-922	0.015
N^6-L-threonylcarbamoyladenosine ^{37}-modified tRNA biosynthesis	PWY0-1587	0.440
nitrate reduction II (assimilatory)	PWY-381	0.197
nitrate reduction V (assimilatory)	PWY-5675	0.211
octanoyl-[acyl-carrier protein] biosynthesis (mitochondria, yeast)	PWY-7388	0.146
oleate biosynthesis III (cyanobacteria)	PWY-7587	0.267
oxygenic photosynthesis	PHOTOALL-PWY	
palmitate biosynthesis II (bacteria and plants)	PWY-5971	0.132
palmitoleate biosynthesis I (from (5Z)-dodec-5-enoate)	PWY-6282	0.278
palmitoleate biosynthesis III (cyanobacteria)	PWY-7589	0.200
pentose phosphate pathway	PENTOSE-P-PWY	
pentose phosphate pathway (non-oxidative branch)	NONOXIPENT-PWY	0.250
pentose phosphate pathway (oxidative branch)	OXIDATIVEPENT-PWY	0.094
pentose phosphate pathway (partial)	P21-PWY	0.111
phosphate acquisition	PWY-6348	0.880
phosphatidylglycerol biosynthesis II (non-plastidic)	PWY4FS-8	
phosphatidylinositol biosynthesis II (eukaryotes)	PWY-7625	0.322
phospholipid remodeling (phosphatidate, yeast)	PWY-7417	0.440

phospholipid remodeling (phosphatidylethanolamine, yeast)	PWY-7409	0.220
photorespiration	PWY-181	0.253
photosynthesis light reactions	PWY-101	0.106
protein ubiquitylation	PWY-7511	0.485
PRPP biosynthesis I	PWY0-662	0.880
pseudouridine degradation	PWY-6019	0.440
putrescine biosynthesis III	PWY-46	0.423
pyrimidine deoxyribonucleotides *de novo* biosynthesis I	PWY-7184	0.302
pyruvate decarboxylation to acetyl CoA	PYRUVDEHYD-PWY	0.606
reactive oxygen species degradation	DETOX1-PWY-1	
ricinoleate biosynthesis	PWY-6429	0.200
S-adenosyl-L-methionine biosynthesis	SAM-PWY	0.119
spermidine biosynthesis I	BSUBPOLYAMSYN-PWY	0.601
stearate biosynthesis II (bacteria and plants)	PWY-5989	0.165
sulfate activation for sulfonation	PWY-5340	0.566
superoxide radicals degradation	DETOX1-PWY	0.212
superpathway of adenosine nucleotides *de novo* biosynthesis I	PWY-7229	
superpathway of L-serine and glycine biosynthesis I	SER-GLYSYN-PWY	
superpathway of phosphatidate biosynthesis (yeast)	PWY-7411	
superpathway of pyrimidine deoxyribonucleotides *de novo* biosynthesis	PWY-7211	
superpathway of pyrimidine ribonucleotides *de novo* biosynthesis	PWY0-162	
TCA cycle II (plants and fungi)	PWY-5690	0.051
TCA cycle IV (2-oxoglutarate decarboxylase)	P105-PWY	0.111
thioredoxin pathway	THIOREDOX-PWY	0.440
triacylglycerol degradation	LIPAS-PWY	1.000

tRNA charging		TRNA-CHARGING-PWY	0.335	
UDP-*N*-acetyl-D-glucosamine biosynthesis II		UDPNACETYLGALSYN-PWY	0.161	
UDP-D-galactose biosynthesis		PWY-7344	0.251	
UMP biosynthesis		PWY-5686	0.340	
urea cycle		PWY-4984	0.297	
UTP and CTP *de novo* biosynthesis		PWY-7176	0.293	
UTP and CTP dephosphorylation I		PWY-7185	0.352	
very long chain fatty acid biosynthesis I		PWY-5080	0.440	
very long chain fatty acid biosynthesis II		PWY-7036	0.275	

Table 24 : Amino acid composition of *Microchloropsis* sp.

The weight of total amino acids is taken from (Volkman et al., 1993). The concentration of protein was calculated to be 0.23 g protein/g dry weight biomass. The coefficients of amino acids in the biomass formation equation are represented in terms of mmol/g dry weight.

Metabolite	% weight of total amino acids	g/g dry weight	Molar mass (g/mol)	Coefficient
Glycine	5.7	0.013	75.1	0.178
L-alanine	7.3	0.017	89.1	0.192
L-arginine	7.6	0.018	174.2	0.102
L-aspartate	8.3	0.019	133.1	0.146
L-cysteine	0.9	0.002	121.2	0.017
L-lysine	5.7	0.013	146.2	0.091
L-leucine	7.0	0.016	131.2	0.125
L-isoleucine	4.8	0.011	131.2	0.086
L-glutamate	10.4	0.024	147.1	0.165
L-histidine	2.1	0.005	155.2	0.032
L-methionine	2.0	0.005	149.2	0.031

L-phenylalanine	6.2	0.015	165.2	0.088
L-proline	7.4	0.017	115.1	0.150
L-serine	5.4	0.013	105.1	0.120
L-tyrosine	4.6	0.011	181.2	0.059
L-tryptophan	1.4	0.003	204.2	0.016
L-valine	6.0	0.014	117.2	0.120
L-threonine	6.3	0.015	119.1	0.124
Hydroxy-L-proline	0.2	0.000	131.1	0.004
L-ornithine	0.2	0.000	132.2	0.003
Aminobutyric acid	0.5	0.001	103.1	0.012

Table 25: Fatty acid composition of *Microchloropsis* sp.

The weight of total fatty acids is taken from (Volkman et al., 1993). The concentration of fatty acids was calculated to be 0.112 g fatty acid/g dry weight biomass. The coefficients of fatty acids in the biomass formation equation are represented in terms of mmol/g dry weight.

Metabolite	% weight of total FA	g /g dry weight	Molar mass (g/mol)	Coefficient
Tetradecanoyl-CoA	5	0.0056	227.37	0.025
Pentadecanoyl-CoA	0.5	0.00056	241.39	0.002
Palmitoyl-CoA	27.8	0.031136	255.42	0.122
Palmitoleoyl-CoA	31.8	0.035616	253.4	0.141
Cis-hexadec-7-enoyl-CoA	0.1	0.000112	253.4	0.0004
(9Z,12Z)-hexadeca-9,12-dienoyl-CoA	0.1	0.000112	251.39	0.0004
Stearoyl-CoA	1	0.00112	283.47	0.004
Oleoyl-CoA	8.3	0.009296	281.46	0.033
Linoleoyl-CoA	1.5	0.00168	279.44	0.006
Alpha-linolenoyl-CoA	0.2	0.000224	277.43	0.001
Gamma-linolenoyl-CoA	0.4	0.000448	277.43	0.002
Di-homo-gamma-linolenate	0.9	0.001008	305.48	0.003
Arachidonoyl-CoA	4	0.00448	303.46	0.015

Eicosapentaenoyl-CoA	16.1	0.018032	301.45	0.060

Table 26 : Sugar composition of *Microchloropsis* sp.

The weight of total sugar is taken from (Volkman et al., 1993). The concentration of sugar was calculated to be 0.13 g sugar/g dry weight biomass. The coefficients of sugar in the biomass formation equation are represented in terms of mmol/g dry weight.

Metabolite	% weight of total sugar	g /gDW	Molar mass (g/mol)	Coefficient
Arabinose	1.2	0.002	150.1	0.011
Fucose	8.9	0.012	164.2	0.073
Galactose	8.8	0.012	180.2	0.065
Glucose	60.5	0.081	180.2	0.450
Mannose	2.0	0.003	180.2	0.015
Rhamnose	11.2	0.015	164.2	0.091
Ribose	3.9	0.005	150.1	0.035
Xylose	3.4	0.005	150.1	0.030

Table 27: Chlorophyll composition of *Microchloropsis* sp.

The concentration of chlorophyll is taken from (Volkman et al., 1993). The coefficient of chlorophyll in the biomass formation equation is represented in terms of mmol/g dry weight.

Metabolite	g /gDW	Molar mass (g/mol)	Coefficient
Chlorophyll a	0.017	892.5	0.019

Table 28 : DNA and RNA composition of *Microchloropsis* sp.

      The concentration of chlorophyll is taken from (Rebolloso-Fuentes et al., 2001). The coefficient of precursors in the biomass formation equation is represented in terms of mmol/g dry weight.

Metabolite		Molar ratio (% mol)	Molar mass (g/mol)	Fraction (g)	% weight	mg/g dry weight	Coefficient
DNA	dTMP	22.8	320.19	7300.4	22.36	0.78	0.002

	dAMP	22.8	331.22	7551.9	23.13	0.81	0.002
	dGMP	27.2	347.22	9444.4	28.92	1.01	0.003
	dCMP	27.2	307.20	8355.8	25.59	0.90	0.003
	Total			32652.4			
	UMP	22.8	324.18	7391.3	21.75	4.39	0.014
	AMP	22.8	347.22	7916.6	23.30	4.71	0.014
RNA	GMP	27.2	363.22	9879.6	29.08	5.87	0.016
	CMP	27.2	323.20	8790.9	25.87	5.23	0.016
	Total			33978.5			
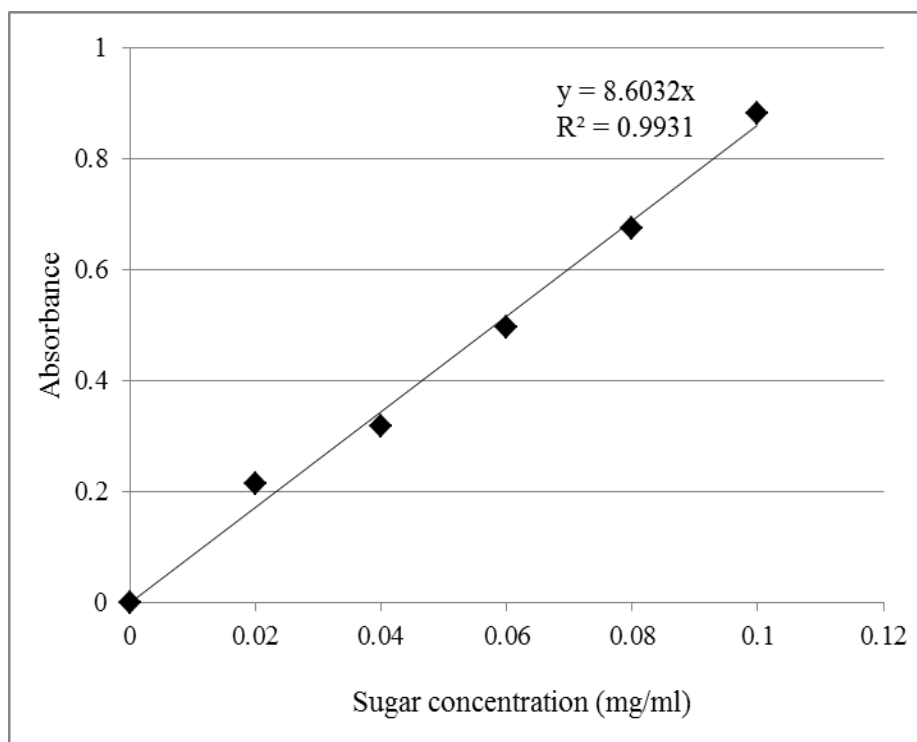
Figure 12: Standard curve for sugar estimation
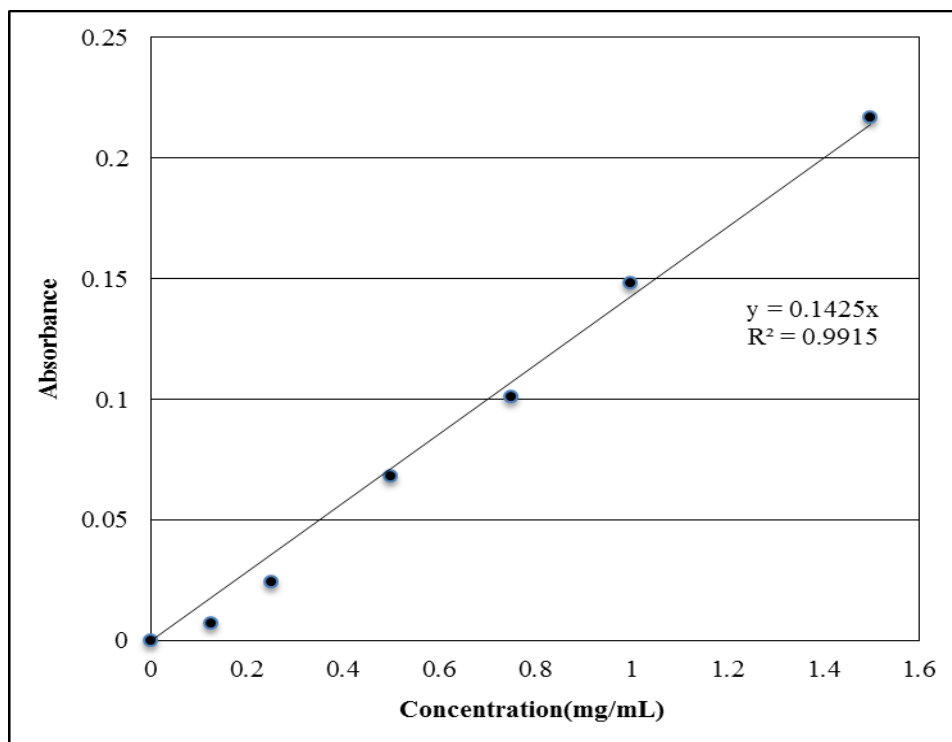
Figure 13 : Standard curve for protein estimation
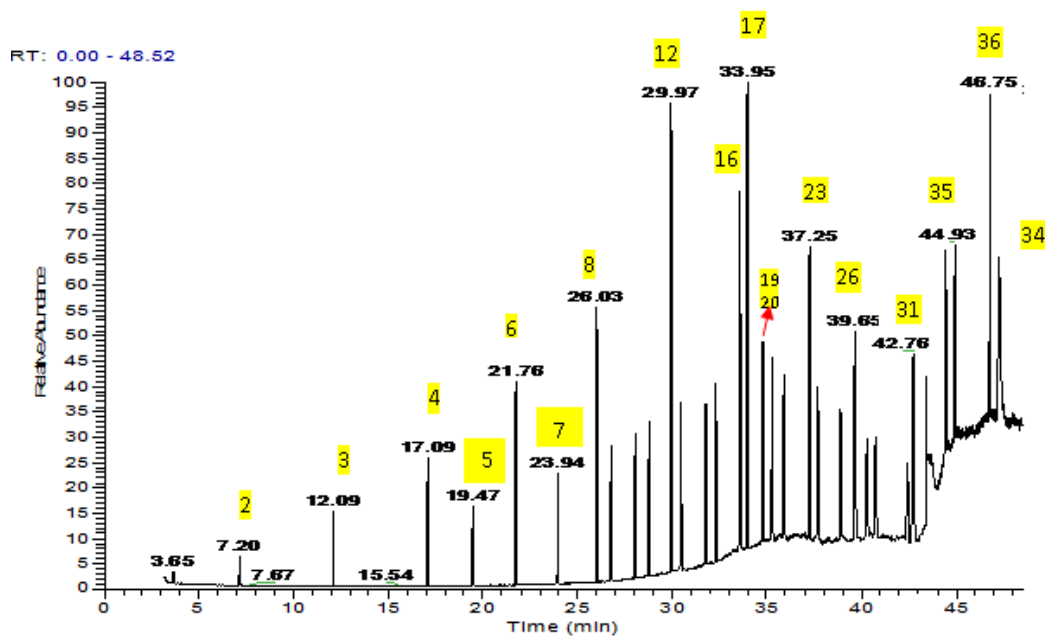
Figure 14 : Chromatogram of FAME mix (the standard). Refer Table 29 for peak Ids.

Table 29 : FAME mix composition and the corresponding peak IDs

Peak ID	FAME components	Peak ID	FAME components
1	C4:0	20	C18:2n6t
2	C6:0	21	C18:3n6
3	C8:0	22	C18:3n3
4	C10:0	23	C20:0
5	C11:0	24	C20:1n9
6	C12:0	25	C20:2
7	C13:0	26	C20:3n6
8	C14:0	27	C20:3n3
9	C14:1	28	C20:4n6
10	C15:0	29	C20:5n3
11	C15:1	30	C21:0
12	C16:0	31	C22:0
13	C16:1	32	C22:1n9
14	C17:0	33	C22:2
15	C17:1	34	C22:6n3
16	C18:0	35	C23:0
17	C18:1n9c	36	C24:0
18	C18:1n9t	37	C24:1n9
19	C18:2n6c		

REFERENCE

Alboresi, A., Perin, G., Vitulo, N., Diretto, G., Block, M. A., Jouhet, J., … Morosinotto, T. (2016). Light Remodels Lipid Biosynthesis in *Nannochloropsis gaditana* by Modulating Carbon Partitioning Between Organelles. *Plant Physiology*, pp.00599.2016. https://doi.org/10.1104/pp.16.00599

Allen, D. K., Libourel, I. G. L., & Shachar‑ Hill, Y. (2009). Metabolic flux analysis in plants: coping with complexity. *Plant, Cell & Environment*, *32*(9), 1241–1257.

Armbrust, E. V., Berges, J. a, Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., … Rokhsar, D. S. (2004). The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science (New York, N.Y.)*, *306*(5693), 79–86. https://doi.org/10.1126/science.1101156

Ben-Amotz, A. (1985). Development of outdoor raceway capable of yielding oil-rich halotolerant microalgae. In *SERl Aquatic Species Review* (pp. 230–244).

Boussiba, S., Vonshak, A., Cohen, Z., Abeliovich, A., Kaplan, D., & Richmond, A. (1985). Development of outdoor system for production of lipid-rich halotolerant microalgae. In *SERl Aquatic Species Review* (pp. 271–288).

Boussiba, S., Vonshak, A., Cohen, Z., Avissar, Y., & Richmond, A. (1987). Lipid and biomass production by the halotolerant microalga *Nannochloropsis salina* . *Biomass*, *12*, 37–47.

Boyle, N. R., & Morgan, J. A. (2009). Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*. *BMC Systems Biology*, *3*, 4. https://doi.org/10.1186/1752-0509-3-4

Burris, J. E. (1981). Effects of oxygen and inorganic carbon concentrations on the photosynthetic quotients of marine-algae. *Marine Biology*, *65*(3), 215–219. https://doi.org/10.1007/bf00397114

Cañavate, J. P., & Lubián, L. M. (1997). Effects of culture age on cryopreservation of marine microalgae. *European Journal of Phycology*, *32*(1), 87–90. https://doi.org/10.1080/09541449710001719405

Cañavate, J. P., & Lubińn, L. M. (1995). Some aspects on the cryopreservation of microalgae

used as food for marine species. *Aquaculture*, *136*(3–4), 277–290. https://doi.org/10.1016/0044-8486(95)01056-4

Caspi, R., Foerster, H., Ingraham, J., Hopkinson, R., Fulcher, C. A., Kaipa, P., … Karp, P. D. (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, *34*(90001), D511–D516. https://doi.org/10.1093/nar/gkj128

Cheirsilp, B., & Torpee, S. (2012). Enhanced growth and lipid production of microalgae under mixotrophic culture condition: Effect of light intensity, glucose concentration and fed-batch cultivation. *Bioresource Technology*, *110*, 510–516. https://doi.org/10.1016/j.biortech.2012.01.125

Chowdhury, A., Zomorrodi, A. R., & Maranas, C. D. (2014). k-OptForce: Integrating Kinetics with Flux Balance Analysis for Strain Design. *PLoS Computational Biology*, *10*(2). https://doi.org/10.1371/journal.pcbi.1003487

Cock, J. M., & Coelho, S. M. (2011). Algal models in plant biology. *Journal of Experimental Botany*, *62*(8), 2425–2430. https://doi.org/10.1093/jxb/err117

Contador, C. a, Rizk, M. L., Asenjo, J. a, & Liao, J. C. (2009). Ensemble modeling for strain development of L-lysine-producing *Escherichia coli*. *Metabolic Engineering*, *11*(4–5), 221–233. https://doi.org/10.1016/j.ymben.2009.04.002

Converti, A., Casazza, A. a., Ortiz, E. Y., Perego, P., & Del Borghi, M. (2009). Effect of temperature and nitrogen concentration on the growth and lipid content of *Nannochloropsis oculata* and *Chlorella vulgaris* for biodiesel production. *Chemical Engineering and Processing: Process Intensification*, *48*(6), 1146–1151. https://doi.org/10.1016/j.cep.2009.03.006

Corteggiani Carpinelli, E., Telatin, A., Vitulo, N., Forcato, C., D'Angelo, M., Schiavon, R., … Valle, G. (2014). Chromosome scale genome assembly and transcriptome profiling of *Nannochloropsis gaditana* in nitrogen depletion. *Molecular Plant*, *7*(2), 323–335. https://doi.org/10.1093/mp/sst120

Dal'Molin, C. G. D. O., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). C4GEM, a genome-scale metabolic model to study C4 plant metabolism. *Plant*

Physiology, *154*(4), 1871–1885. https://doi.org/10.1104/pp.110.166488

Dal'Molin, C. G. D. O., Quek, L.-E., Palfreyman, R. W., Nielsen, L. K., Gomes de Oliveira Dal'Molin, C., Quek, L.-E., … Nielsen, L. K. (2011). AlgaGEM--a genome-scale metabolic reconstruction of algae based on the *Chlamydomonas reinhardtii* genome. *BMC Genomics*, *12 Suppl 4*(Suppl 4), S5. https://doi.org/10.1186/1471-2164-12-S4-S5

Das, P., Lei, W., Aziz, S. S., & Obbard, J. P. (2011). Enhanced algae growth in both phototrophic and mixotrophic culture under blue light. *Bioresource Technology*, *102*(4), 3883–3887. https://doi.org/10.1016/j.biortech.2010.11.102

de la Vega, M., Díaz, E., Vila, M., & León, R. (2011). Isolation of a new strain of *Picochlorum* sp and characterization of its potential biotechnological applications. *Biotechnology Progress*, *27*(6), 1535–1543. https://doi.org/10.1002/btpr.686

de Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiology*, *152*(2), 579–589. https://doi.org/10.1104/pp.109.148817

Doyle, M. A., MacRae, J. I., De Souza, D. P., Saunders, E. C., McConville, M. J., & Likić, V. A. (2009). LeishCyc: a biochemical pathways database for *Leishmania major*. *BMC Systems Biology*, *3*(1), 57. https://doi.org/10.1186/1752-0509-3-57

Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Intercoder reliability and validity of webplotdigitizer in extracting graphed data. *Behavior Modification*, *41*(2), 323–339. https://doi.org/10.1177/0145445516673998

DuBois, M., Gilles, K. a., Hamilton, J. K., Rebers, P. a., & Smith, F. (1956). Colorimetric Method for Determination of Sugars and Related Substances. *Analytical Chemistry*, *28*(3), 350–356. https://doi.org/10.1021/ac60111a017

Everest, S. A., Hipkin, C. R., & Syrett, P. J. (1986). Enzyme activities in some marine phytoplankters and the effect of nitrogen limitation on nitrogen and carbon metabolism in *Chlorella stigmatophora*. *Marine Biology*, *90*(2), 165–172.

Fang, X., Wei, C., Zhao-Ling, C., & Fan, O. (2004). Effects of organic carbon sources on cell growth and eicosapentaenoic acid content of *Nannochloropsis sp. Journal of Applied*

Phycology, *16*(2004), 499–503. https://doi.org/10.1007/s10811-005-5520-9

Fawley, M. W., Jameson, I., & Fawley, K. P. (2015). The phylogeny of the genus *Nannochloropsis* (Monodopsidaceae, Eustigmatophyceae), with descriptions of *N. australis sp. nov* . and *Microchloropsis gen. nov* . *Phycologia*, *54*(5), 545–552. https://doi.org/10.2216/15-60.1

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., & Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nature Reviews. Microbiology*, *7*(2), 129–143. https://doi.org/10.1038/nrmicro1949

Galloway, R. E. (1990). Selective conditions and isolation of mutants in salt-tolerant, lipid-producing microalgae. *Journal of Phycology*, *26*(4), 752–760.

Garcés, R., & Mancha, M. (1993). One-step lipid extraction and fatty acid methyl esters preparation from fresh plant tissues. *Analytical Biochemistry*, *211*(1), 139–143. https://doi.org/10.1006/abio.1993.1244

Green, M. L., & Karp, P. D. (2004). A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, *5*, 76. https://doi.org/10.1186/1471-2105-5-76

Guillard, R. R. . (1975). Culture of phytoplankton for feeding marine invertebrates. In *Culture of marine invertebrate animals* (pp. 29–60). Springer US.

H. Ryther, J., & Guillard, R. R. . (1962). Studies of marine planktonic diatoms: I. *Cyclotella Nana* Hustedt, and *Detonula Confervacea* (CLEVE) Gran. *Canadian Journal of Microbiology*, *8*(1140), 229–239.

Hibberd, D. J. (1981). Notes on the taxonomy and nomenclature of the algal classes Eustigmatophyceae and Tribophyceae (synonym Xanthophyceae). *Botanical Journal of the Linnean Society*, *82*(2), 93–119. https://doi.org/10.1111/j.1095-8339.1981.tb00954.x

Hibberd, D. J., & Leedale, G. F. (1970). Eustigmatophyceae - A new algal class with unique organization of the motile cell. *Nature*, *225*(5234), 758–760. https://doi.org/10.1038/225758b0

Hirayama, K., Maruyama, I., & Maeda, T. (1989). Nutritional effect of freshwater *Chlorella* on growth of the rotifer *Brachionus plicatilis*. *Hydrobiologia*, *186–187*(1), 39–42. https://doi.org/10.1007/BF00048894

Hockin, N. L., Mock, T., Mulholland, F., Kopriva, S., & Malin, G. (2012). The Response of Diatom Central Carbon Metabolism to Nitrogen Starvation Is Different from That of Green Algae and Higher Plants. *Plant Physiology*, *158*(1), 299–312. https://doi.org/10.1104/pp.111.184333

Huertas, I. E., Espie, G. S., Colman, B., & Lubian, L. M. (2000). Light-dependent bicarbonate uptake and CO2 efflux in the marine microalga *Nannochloropsis gaditana*. *Planta*, *211*(1), 43–49. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10923702

Huertas, I. E., & Lubian, L. M. (1998). Comparative study of dissolved inorganic carbon utilization and photosynthetic responses in *Nannochloris (Chlorophyceae)* and *Nannochloropsis* (Eustigmatophyceae) species. *Canadian Journal of Botany-Revue Canadienne De Botanique*, *76*, 1104–1108. https://doi.org/10.1139/cjb-76-6-1104

Hyduke, D., Schellenberger, J., Que, R., Fleming, R., Thiele, I., Orth, J., … Palsson, B. Ø. (2011). COBRA Toolbox 2 . 0. *Protocol Exchange*, *22*.

Jacobsen, J. H., Rosgaard, L., Sakuragi, Y., & Frigaard, N.-U. (2011). One-step plasmid construction for generation of knock-out mutants in cyanobacteria: studies of glycogen metabolism in *Synechococcus* sp. PCC 7002. *Photosynthesis Research*, *107*(2), 215–221. https://doi.org/10.1007/s11120-010-9613-1

James, C. M., & Al-Khars, A. M. (1990). An intensive continuous culture system using tubular photobioreactors for producing microalgae. *Aquaculture*, *87*(3–4), 381–393. https://doi.org/10.1016/0044-8486(90)90075-X

Jiang, Y., Yoshida, T., & Quigg, A. (2012). Photosynthetic performance, lipid production and biomass composition in response to nitrogen limitation in marine microalgae. *Plant Physiology and Biochemistry*, *54*, 70–77. https://doi.org/10.1016/j.plaphy.2012.02.012

Jinkerson, R. E., Radakovits, R., & Posewitz, M. C. (2013). Genomic insights from the oleaginous model alga, (February), 37–43.

Karp, P. D. (2002). The EcoCyc Database. *Nucleic Acids Research*, *30*(1), 56–58. https://doi.org/10.1093/nar/30.1.56

Karp, P. D., Paley, S., & Romero, P. (2002). The Pathway Tools software. *Bioinformatics*, *18*(Suppl 1), S225–S232. https://doi.org/10.1093/bioinformatics/18.suppl_1.S225

Karp, Peter D, Latendresse, M., & Caspi, R. (2011). The pathway tools pathway prediction algorithm. *Standards in Genomic Sciences*, *5*(3), 424–429. https://doi.org/10.4056/sigs.1794338

Karp, Peter D, Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., … Caspi, R. (2010). Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, *11*(1), 40–79. https://doi.org/10.1093/bib/bbp043

Kilian, O., Benemann, C. S. E., Niyogi, K. K., & Vick, B. (2011). High-efficiency homologous recombination in the oil-producing alga *Nannochloropsis sp. Proceedings of the National Academy of Sciences of the United States of America*, *108*(52), 21265–21269. https://doi.org/10.1073/pnas.1105861108

Koven, W. M., Tandler, A., Kissil, G. W., Sklan, D., Friezlander, O., & Harel, M. (1990). The effect of dietary (n-3) polyunsaturated fatty acids on growth, survival and swim bladder development in *Sparus aurata* larvae. *Aquaculture*, *91*(1–2), 131–141. https://doi.org/10.1016/0044-8486(90)90182-M

Lee, S. K., Chou, H., Ham, T. S., Lee, T. S., & Keasling, J. D. (2008). Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current Opinion in Biotechnology*, *19*(6), 556–563. https://doi.org/10.1016/j.copbio.2008.10.014

Li, J., Han, D., Wang, D., Ning, K., Jia, J., Wei, L., … Xu, J. (2014). Choreography of Transcriptomes and Lipidomes of *Nannochloropsis* Reveals the Mechanisms of Oil Synthesis in Microalgae. *The Plant Cell*, *26*(4), 1645–1665. https://doi.org/10.1105/tpc.113.121418

Liu, B., Vieler, A., Li, C., Jones, a D., & Benning, C. (2013). Triacylglycerol profiling of microalgae *Chlamydomonas reinhardtii* and *Nannochloropsis oceanica*. *Bioresource*

Technology, *146*, 310–316. https://doi.org/10.1016/j.biortech.2013.07.088

Loira, N., Mendoza, S., Paz Cortés, M., Rojas, N., Travisany, D., Genova, A. Di, … Maass, A. (2017). Reconstruction of the microalga *Nannochloropsis salina* genome-scale metabolic model with applications to lipid production. *BMC Systems Biology*, *11*(1), 66. https://doi.org/10.1186/s12918-017-0441-1

Lopes, V. R., & Vasconcelos, V. M. (2011). Bioactivity of benthic and picoplanktonic estuarine cyanobacteria on growth of photoautotrophs: inhibition versus stimulation. *Marine Drugs*, *9*(5), 790–802. https://doi.org/10.3390/md9050790

Lubian, L. M. (1982). *Nannochloropsis gaditana sp.* nov., una nueva Eustigmatophyceae marina. *Lazaroa*, *4*, 287–293. https://doi.org/-

Mackie, A., Keseler, I. M., Nolan, L., Karp, P. D., & Paulsen, I. T. (2013). Dead end metabolites - Defining the known unknowns of the *E. coli* metabolic network. *PloS One*, *8*(9), e75210. https://doi.org/10.1371/journal.pone.0075210

Marudhupandi, T., Sathishkumar, R., & Kumar, T. T. A. (2016). Heterotrophic cultivation of *Nannochloropsis salina* for enhancing biomass and lipid production. *Biotechnology Reports*, *10*, 8–16. https://doi.org/10.1016/j.btre.2016.02.001

May, P., Christian, J.-O., Kempa, S., & Walther, D. (2009). ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii. BMC Genomics*, *10*(1), 209. https://doi.org/10.1186/1471-2164-10-209

McGinnis, S., & Madden, T. L. (2004). BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, *32*(WEB SERVER ISS.), 20–25. https://doi.org/10.1093/nar/gkh435

Mitra, M., Patidar, S. K., George, B., Shah, F., & Mishra, S. (2015). A euryhaline *Nannochloropsis gaditana* with potential for nutraceutical (EPA) and biodiesel production. *Algal Research*, *8*, 161–167. https://doi.org/10.1016/j.algal.2015.02.006

Mitra, M., Patidar, S. K., & Mishra, S. (2015). Integrated process of two stage cultivation of *Nannochloropsis* sp. for nutraceutically valuable eicosapentaenoic acid along with biodiesel. *Bioresource Technology*, *193*, 363–369.

https://doi.org/10.1016/j.biortech.2015.06.033

Mourente, G., Lubian, L. M., & Odriozola, J. M. (1989). Total lipid fatty acid composition as taxonomic index of some marine microalgae used as food for marine cultures, 147–154.

Mueller, L. A., Zhang, P., & Rhee, S. Y. (2003). AraCyc: A Biochemical Pathway Database for *Arabidopsis*. *Plant Physiology*, *132*(2), 453–460. https://doi.org/10.1104/pp.102.017236

Munoz, J., & Merrett, M. J. (1989). Inorganic-carbon transport in some marine eukaryotic microalgae. *Planta*, *178*(4), 450–455. https://doi.org/10.1007/BF00963814

Neidhardt, F. C., Ingraham, J. L., & Schaechter., M. (1990). *Physiology of the bacterial cell: a molecular approach.* Sunderland, MA: Sinauer Associates.

Nitsan, Z., Mokady, S., & Sukenik, A. (1999). Enrichment of poultry products with fatty acids by dietary supplementation with the alga *Nannochloropsis* and mantur oil. *Journal of Agricultural and Food Chemistry*, *47*(12), 5127–5132. https://doi.org/10.1021/jf981361p

Oberhardt, M. a, Palsson, B. Ø., & Papin, J. a. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, *5*(320), 320. https://doi.org/10.1038/msb.2009.77

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–248. https://doi.org/10.1038/nbt.1614

Owens, T. G., Gallagher, J. C., & Alberte, R. S. (1987). Photosynthetic light-harvesting function of violaxanthin in *Nannochloropsis* spp. (Eustigmatophyceae). *Journal of Phycology*. https://doi.org/10.1111/j.0022-3646.1987.00079.x

Radakovits, R., Eduafo, P. M., & Posewitz, M. C. (2011). Genetic engineering of fatty acid chain length in *Phaeodactylum tricornutum*. *Metabolic Engineering*, *13*(1), 89–95. https://doi.org/10.1016/j.ymben.2010.10.003

Radakovits, R., Jinkerson, R. E., Fuerstenberg, S. I., Tae, H., Settlage, R. E., Boore, J. L., & Posewitz, M. C. (2012). Draft genome sequence and genetic transformation of the oleaginous alga *Nannochloropis gaditana*. *Nature Communications*, *3*, 686. https://doi.org/10.1038/ncomms1688

Raso, S., van Genugten, B., Vermuë, M., & Wijffels, R. H. (2012). Effect of oxygen concentration on the growth of *Nannochloropsis sp.* at low light intensity. *Journal of Applied Phycology*, *24*(4), 863–871. https://doi.org/10.1007/s10811-011-9706-z

Rebolloso-Fuentes, M. M., Navarro-Pérez, A., García-Camacho, F., Ramos-Miras, J. J., & Guil-Guerrero, J. L. (2001). Biomass nutrient profiles of the microalga *Nannochloropsis*. *Journal of Agricultural and Food Chemistry*, *49*(6), 2966–2972. https://doi.org/10.1021/jf0010376

Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., … Palsson, B. Ø. (2006). Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(46), 17480–17484. https://doi.org/10.1073/pnas.0603364103

Renaud, S. M., Parry, D. L., Thinh, L. Van, Kuo, C., Padovan, A., & Sammy, N. (1991). Effect of light intensity on the proximate biochemical and fatty acid composition of *Isochrysis sp.* and *Nannochloropsis oculata* for use in tropical aquaculture. *Journal of Applied Phycology*, *3*(1), 43–53. https://doi.org/10.1007/BF00003918

Saha, R., Chowdhury, A., & Maranas, C. D. (2014). Recent advances in the reconstruction of metabolic models and integration of omics data. *Current Opinion in Biotechnology*, *29*(1), 39–45. https://doi.org/10.1016/j.copbio.2014.02.011

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., … Palsson, B. Ø. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature Protocols*, *6*(9), 1290–1307. https://doi.org/10.1038/nprot.2011.308

Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., … Rhee, S. Y. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology*, *173*(4), 2041–2059. https://doi.org/10.1104/pp.16.01942

Shah, A. R., Ahmad, A., Srivastava, S., & Jaffar Ali, B. M. (2017). Reconstruction and analysis of a genome-scale metabolic model of *Nannochloropsis gaditana*. *Algal Research*, (July). https://doi.org/10.1016/j.algal.2017.08.014

Shastri, A. a., & Morgan, J. a. (2005). Flux balance analysis of photoautotrophic metabolism.

Biotechnology Progress, *21*(6), 1617–1626. https://doi.org/10.1021/bp050246d

Simionato, D., Block, M. a, La Rocca, N., Jouhet, J., Maréchal, E., Finazzi, G., & Morosinotto, T. (2013). The response of *Nannochloropsis gaditana* to nitrogen starvation includes de novo biosynthesis of triacylglycerols, a decrease of chloroplast galactolipids, and reorganization of the photosynthetic apparatus. *Eukaryotic Cell*, *12*(5), 665–676. https://doi.org/10.1128/EC.00363-12

Simionato, D., Sforza, E., Corteggiani Carpinelli, E., Bertucco, A., Giacometti, G. M., & Morosinotto, T. (2011). Acclimation of *Nannochloropsis gaditana* to different illumination regimes: effects on lipids accumulation. *Bioresource Technology*, *102*(10), 6026–6032. https://doi.org/10.1016/j.biortech.2011.02.100

Stein, L. (2001). Genome Annotation: From sequence to biology. *Nature Reviews Genetics*, *2*, 493–503.

Stephanopoulos, G., Aristidou, A., & Nielsen, J. (1999). *Metabolic Engineering: Principles and Methodologies*. Elsevier Science.

Suen, Y., Hubbard, J. S., Holzer, G., & Tornabene, T. G. (1987). Total lipid production of the green alga *Nannochloropsis sp.* QII under different nitrogen regimes. *Journal of Phycology*, *23*, 289–296. https://doi.org/10.1111/j.1529-8817.1987.tb04137.x

Sukenik, A. (1991). Ecophysiological considerations in the optimization of eicosapentaenoic acid production by *Nannochloropsis sp.*(Eustigmatophyceae). *Bioresource Technology*, *35*(3), 263–269. https://doi.org/10.1016/0960-8524(91)90123-2

Sukenik, A., Carmeli, Y., & Berner, T. (1989). Regulation of fatty acid composition by irradiance level in the eustigmatophyte *Nannochloropsis sp. Journal of Phycology*, *25*(4), 686–692.

Sukenik, A., Tchernov, D., Kaplan, A., Huertas, I. E., Lubian, L. M., & Livne, A. (1997). Uptake, efflux, and photosynthetic utilization of inorganic carbon by the marine Eustigmatophyte *Nannochloropsis* sp.1. *Journal of Phycology*, *33*(6), 969–974. https://doi.org/10.1111/j.0022-3646.1997.00969.x

Sukenik, A., Zmora, O., & Carmeli, Y. (1993). Biochemical quality of marine unicellular algae

with special emphasis on lipid composition. II. *Nannochloropsis sp. Aquaculture*, *117*(3–4), 313–326. https://doi.org/10.1016/0044-8486(93)90328-V

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), 93–121. https://doi.org/10.1038/nprot.2009.203

Urbanczyk-Wochniak, E., & Sumner, L. W. (2007). MedicCyc: A biochemical pathway database for *Medicago truncatula*. *Bioinformatics*, *23*(11), 1418–1423. https://doi.org/10.1093/bioinformatics/btm040

Van Moerkercke, A., Fabris, M., Pollier, J., Baart, G. J. E., Rombauts, S., Hasnain, G., … Goossens, A. (2013). CathaCyc, a metabolic pathway database built from *Catharanthus roseus* RNA-seq data. *Plant and Cell Physiology*, *54*(5), 673–685. https://doi.org/10.1093/pcp/pct039

Vieler, A., Wu, G., Tsai, C.-H., Bullard, B., Cornish, A. J., Harvey, C., … Benning, C. (2012). Genome, functional gene annotation, and nuclear transformation of the heterokont oleaginous alga *Nannochloropsis oceanica CCMP1779*. *PLoS Genetics*, *8*(11), e1003064. https://doi.org/10.1371/journal.pgen.1003064

Volkman, J. K., Brown, M. R., Dunstan, G. a, & Jeffrey, S. W. (1993). The biochemical-composition of marine microalgae from the class Eustigmatophyceae. *Journal of Phycology*, *29*(October), 69–78. https://doi.org/10.1111/j.1529-8817.1993.tb00281.x

Wang, D., Ning, K., Li, J., Hu, J., Han, D., Wang, H., … Xu, J. (2014). *Nannochloropsis* genomes reveal evolution of microalgal oleaginous traits. *PLoS Genetics*, *10*(1), e1004094. https://doi.org/10.1371/journal.pgen.1004094

Weeks, D. P. (2011). Homologous recombination in *Nannochloropsis*: a powerful tool in an industrially relevant alga. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(52), 20859–20860. https://doi.org/10.1073/pnas.1118670109

Werman, M. J., Sukenik, A., & Mokady, S. (2003). Effects of the marine unicellular alga *Nannochloropsis* sp. to reduce the plasma and liver cholesterol levels in male mats fed on diets with cholesterol. *Bioscience, Biotechnology, and Biochemistry*, *67*(10), 2266–2268.

https://doi.org/10.1271/bbb.67.2266

Xu, F., Cai, Z. L., Cong, W., & Ouyang, F. (2004). Growth and fatty acid composition of *Nannochloropsis sp.* grown mixotrophically in fed-batch culture. *Biotechnology Letters*, *26*(17), 1319–1322. https://doi.org/10.1023/B:BILE.0000045626.38354.1a

Yee, W., Ahmad, A., & Cha, T. S. (2012). Factors affecting *Agrobacterium*-mediated genetic transformation of marine microalga, *Nannochloropsis* sp. *Journal of Sustainability Science and Management*, *7*(2), 153–163.

Yoshikawa, K., Kojima, Y., Nakajima, T., Furusawa, C., Hirasawa, T., & Shimizu, H. (2011). Reconstruction and verification of a genome-scale metabolic model for *Synechocystis* sp. PCC6803. *Applied Microbiology and Biotechnology*, *92*(2), 347–358. https://doi.org/10.1007/s00253-011-3559-x

Zamboni, N. (2011). 13C metabolic flux analysis in complex systems. *Current Opinion in Biotechnology*, *22*(1), 103–108. https://doi.org/10.1016/j.copbio.2010.08.009

Zhang, P., Dreher, K., Karthikeyan, A., Chi, A., Pujar, A., Caspi, R., … Rhee, S. Y. (2010). Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiology*, *153*(4), 1479–1491. https://doi.org/10.1104/pp.110.157396

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, *7*(12), 203–214. https://doi.org/10.1089/10665270050081478

Zhu, C. J., & Lee, Y. K. (1997). Determination of biomass dry weight of marine microalgae. *Journal of Applied Phycology*, *9*(2), 189–194. https://doi.org/10.1023/A:1007914806640

Publications

1. Akhila George, John Beardall, Madhu Chetty, Santanu Dasgupta, & Pramod P.Wangikar. MgdCyc: A biochemical pathway database for *Microchloropsis gaditana CCMP526*. (To be submitted).

2. Akhila George, John Beardall, Madhu Chetty, Santanu Dasgupta, & Pramod P.Wangikar. Development and Analysis of genome-scale metabolic reconstruction of *Microchloropsis gaditana CCMP526.* (To be submitted).

Acknowledgement

I take this opportunity to convey my gratitude to everyone who encouraged and supported me throughout my doctoral studies. I wish to express my heartfelt gratitude my supervisor, **Prof. Pramod. P. Wangikar,** Indian Institute of Technology, Bombay and co-supervisor, **Prof. John Beardall**, Monash University, for their guidance and motivation. I am also obliged to my co-supervisors, **Prof. Madhu Chetty** and **Dr. Santanu Dasgupta**, Reliance Industries Limited, for their encouragement and guidance. I thank **Dr. Tomal Dattaroy and Dr. Uma Shankar** from Reliance Industries Limited, for their constant support. I thank **Prof. Murali Sastry** and express special thanks to all the staff members of The IITB-Monash Research Academy for their never-failing support throughout this course. I acknowledge **Reliance Industries Limited** for funding my PhD project.

I am thankful to **Prof. Yogendra Shastri**, IIT Bombay and **Prof. Sureshkumar Balasubramanian**, Monash University for being my RPC members and for their valuable suggestions and feedback. I specially thank **Dr. S. Krishnakumar, Dr. Pramod R. Somvanshi, Dr. Ajay Nair**, **Dr. Charulata Prasannan and Dr. John Irudayaraj** for sharing their experience, knowledge and skills with me. I express my appreciation and gratitude to my lab mates in Biosystem Engineering Lab, IIT Bombay who have been supporting and encouraging me during the course. I also thank lab mates in John's lab, Monash University who patiently explained techniques and helped me to settle.

I gratefully acknowledge the support of my close friends, **Alice Mathew, Anoop George, Arun Mathew, Gisha Mathew, Jaison Davis** and **P. Krishnendu**, for always being there through my ups and downs. I express my special thanks to the friends from "**Kaapisangham**" and "**Koinonia**" group for their generous care and support throughout this journey.

I express my deep gratitude to my parents, grandparents, sister and parents-in law who stood by my side and motivated me to successfully complete my doctoral studies. Last but not least, I thank my beloved husband, **Shon Mathew Thomas**, for his unwavering support and encouragement during my pursuit of PhD degree.