Improved functional annotations are key to realize the potential of algal biotechnology

Maarten J.M.F. Reijnders

Improved functional annotations are key to realize the potential of algal biotechnology

Maarten J.M.F. Reijnders

Thesis committee

Promotor
Prof. Dr Vitor A. P. Martins dos Santos
Professor of Systems and Synthetic Biology
Wageningen University & Research

Promotor
Prof. Dr Gerrit Eggink
Special professor, Industrial Biotechnology
Wageningen University & Research

Co-promotor
Dr Peter J. Schaap
Associate professor, Systems and Synthetic Biology
Wageningen University & Research

Other members
Prof. Dr Robert D. Hall, Wageningen University & Research
Dr Aalt D.J. van Dijk, Wageningen University & Research
Dr Richard A. Notebaart, Wageningen University & Research
Dr Filipe Branco dos Santos, University of Amsterdam

This research was conducted under the auspices of the graduate school VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences)

Improved functional annotations are key to realize the potential of algal biotechnology

Maarten J.M.F. Reijnders

Thesis
submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 13 November 2018
at 1:30 p.m. in the Aula

Maarten J.M.F. Reijnders
Improved functional annotations are key to realize the potential of algal biotechnology,
196 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018).
With references, with summary in English.

ISBN: 978-94-6343-535-2
DOI: 10.18174/462782

Table of contents

Chapter 1

General introduction

1. Biotechnology

Biotechnology is the exploitation of biological resources for the benefit of human society, and has been practiced for thousands of years [3]. The earliest forms of biotechnology were the domestication of animals and crops such as wheat [4]. Later biotechnology extended to the use of microorganisms such as yeast for fermenting beer, wine, and bread [5]. These are examples of using organisms in their natural state to benefit human life. Modern biotechnology, however, tries to understand an organism and its genetics in order for it to be modified and exploited [6]. For example, by manipulating a microorganism its growth conditions, removing genes, or introducing new genes. Manipulations like this can lead to increased growth [7], increased product formation [8, 9], or even the synthesis of a compound the organism was not capable of synthesizing before [8, 10, 11]. A big part in understanding the genetics of organisms for modern biotechnology is the use of computational biology to create *in-silico* driven hypotheses for a wide range of topics including bioprocess optimization for product synthesis, simulating the effects of medicine, whole cell analysis of host organisms, and many more.

1.2 Computational biology for biotechnology

For computational biology there are two important fields: bioinformatics, and systems biology. Bioinformatics is a general term used for many *in silico* methodologies studying for example omics data, protein structures, and cellular organization. This is done using computer programming to handle big data, statistics, and mathematics. Most interesting for biotechnology is the study of omics data, called genomics [12-14]. In genomics, bioinformatics is used to characterize the genome, identify the proteins made, the function of each protein, when and where they are expressed, how they interact, and what biological pathways they are involved in. Systems biology, on the other hand, studies the interaction between the components of a system using mathematical modeling of these systems ranging from a specific biological process of an organism, to a genome scale model, and even how different organisms interact as a system in cases such as symbiosis or parasitism. Together, these two fields of computational biology are able to *in silico* predict what effects environmental factors and gene editing have on an organism. With these predictions obtained through bioinformatics and systems biology

research, it is possible to improve the outcome of wet-lab biotechnological research.

1.3 Artemisinin production: an example of design driven biotechnology

An excellent example of design driven biotechnology is the development of a *Saccharomyces cerevisiae* strain able to produce high amounts of artemisinic acid, a precursor for artemisinin drugs used in the treatment of malaria [8, 15]. In the development of this strain, researchers utilized a wide range of tools to successfully introduce a complete biosynthetic pathway. Using bioinformatics and systems biology techniques they were one of the first to successfully implement a Design-Build-Test-Learn cycle, as described in Figure 1 by Niels and Keasling [16]. This cycle consists of four interacting modules: designing a biological system, building the biological system, testing the biological system, and learning from the biological system. In an ideal case scenario each cycle leads to a better design and eventually an engineered strain that is able to efficiently perform the desired task, such as more synthesis of a product or increased growth.

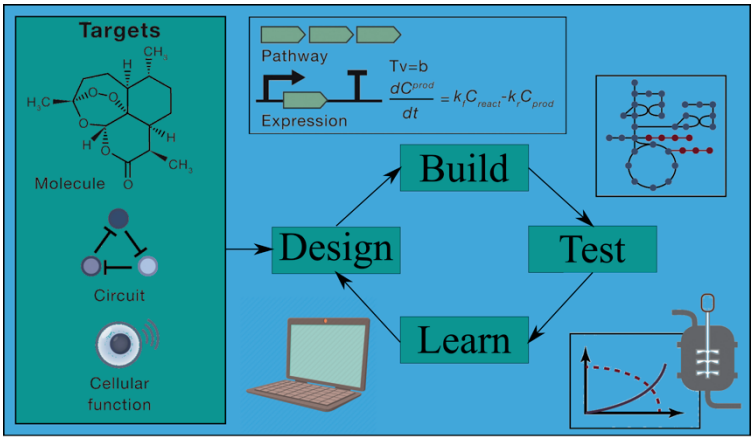


Figure 1: An example design-build-test-learn cycle as described and adapted from Nielsen and Keasling [16]. First a target molecule, regulatory circuit, and host are identified. These are used to design the system. After, the design is built in the wet-lab, and the resulting strain is tested and characterized. Following these tests and characterizations, the learned knowledge is used to improve the design of the system for the next cycle.

The ‘design’ module of the cycle was applied in several forms for engineering of the artemisinic acid *S. cerevisiae* strain. Some of the techniques that were used are comparative genomics between different *S. cerevisiae* strains to identify the cause for the differences in their behaviour, and transcriptomic analysis to identify rate limiting reactions. After many iterations, these researchers managed to make a commercially viable yeast-based artemisinin production pipeline [17]. This example of artemisinin production in yeast thus utilized computational biology to overcome initial challenges and improve their yield. This resulted in dozens of updated yeast strains based on their design-driven hypotheses [8, 17]. Computational predictions were done on specific components, but also on how these components would fit in the yeast its metabolism. With this approach, researchers managed to engineer a yeast strain producing the product they want, in commercially viable amounts, and minimizing metabolic side effects.

Other examples of successful design-driven, or computational design-driven, metabolic engineering and synthetic biology is the production of 1,3-propanediol by DuPont using *Escherichia coli*, isobutanol by Gevo and Butamax using yeast, and 1,4-butanediol Genomatica using *Escherichia coli* [18].

2. Untapped potential: microalgae for biotechnology

2.1 Microalgae and their products

With the above example in mind it is interesting to apply a similar strategy to microalgae. Microalgae have a lot of biotechnological potential, particularly those that are informally called green microalgae [19-22], and are often capable of producing large amounts of lipids, sugars, or other valuable metabolites [23-27]. For example, many microalgae are oleaginous, meaning their biomass can consist of 40% or more of high-value lipids like triacylglycerol [28, 29]. These lipids can be used to produce biodiesel, bioplastics, pharmaceuticals, and many more [30-32]. Other biotechnology applications for microalgae focus on its value in food, nutrition, and cosmetics [33, 34]. The biomass of many microalgae consist of a high amount of proteins and other nutrients, which also makes them a cheap addition to food sources [35]. In the case of animal feed, microalgae are often used as feedstock for aquatic species. When it comes to human consumption, some microalgal species showed to be beneficial to human health in a variety of ways, such as a source for omega-3 fatty acids [36], anti-allergy benefits [37], and anti-viral benefits [38].

In recent history research and commercial activities are more focused towards their potential for producing high-value compounds such as lipids and sugars that can be used in commercial products [39]. However, in this case, commercial use of microalgae is only feasible if it outperforms existing sources such as fossil fuels, chemical production of compounds, or various higher order plants. Production of many interesting compounds often comes at a high metabolic cost, resulting in slow growth and other diminishing effects [29]. Biotechnological research for microalgae aims to optimize the product formation of microalgae while minimizing these diminishing effects.

2.2 Current constraints for microalgal biotechnology

Currently, when it comes to finished and commercialized products, microalgal biotechnology for the production of high-value compounds is only available for a select few companies, often not further developed beyond pilot scales [40]. The general state of microalgal genetics is on a proof-of-principle level. On one hand the techniques for microalgal genome editing are still in development or coming of age [41], and on the other hand there is an inherent lack of genetic knowledge of microalgae. A key example is that of biofuels. Biofuel has been one of the main selling points for microalgal research over the

past decade. However, recent studies suggest that in their current state microalgae for biofuels will not be commercially viable [42, 43]. A lot of research regarding microalgae and biofuels has been done on the bioprocess engineering level, for example optimizing bioreactors and growth conditions for microalgae by maximizing their light uptake or nutrient availability [44, 45]. However, only optimizing microalgae in this manner will not be commercially viable for the foreseeable future due to metabolic constraints [43]. One of the main metabolic constraints is the photosynthetic limitation of microalgae. The most optimistic estimates in which sunlight is converted to energy are up to 8% [42]. On top of that, the actual conversion of this photosynthetic energy to biomass is as low as 35% [46]. This means that a huge amount of photosynthetic energy is used in the intermediate metabolic processes of microalgae. These limitations are the main reason why microalgae are not economically viable for biofuel in the foreseeable future. Genomics research will be needed to study if there are any fixable metabolic constraints, and if so, genetic engineering will be needed to make microalgae more efficient in their metabolic processes.

In recent years microalgal research has been moving towards the understanding of their genomes and metabolic processes. For example, many efforts are made towards identifying genome-wide differential expression during starch and lipid production, and during nitrogen limitation or starvation. Another example is the aim to identify molecular switches between starch and lipid production, and how these are regulated [47]. In a more futuristic approach, some projects are even attempting to engineer a more efficient photosynthesis [48, 49]. However, what all these have in common is that they are at a fundamental level research that has yet to be implemented for production purposes. This is where synthetic biology can play a role.

2.3 Microalgal genomics and synthetic biology

Synthetic biology is the systematic characterization and usage of standardized genetic parts, and the engineering of these parts into organisms to create new or more products. Applying synthetic biology principles to microalgae can result in microalgal cell factories, able to synthesize various products at a high rate and commercially viable cost. However, synthetic biology is heavily reliant on precise functional knowledge, and despite recent efforts a genomic understanding of microalgae still lags behind in comparison to that of many other species. For example, the genome of *Saccharomyces cerevisiae* got published in 1996 [50], that of *Escherichia coli* in 1997 [51], and that of

Arabidopsis thaliana in 2000 [52]. In comparison, the first published genome of green algae is *Ostreococcus taurii* in 2006 [53], and the model species *Chlamydomonas reinhardtii* in 2007 [54]. If one compares these dates to those of the aforementioned species it is extremely late and exemplifies the general lack of research on microalgal genomics during the early omics age. Current next generation sequencing technologies make up for the late start of genome sequencing for microalgae, but due to the late start, so far little efforts have been made to experimentally characterize specific algal genomic features. Characterizing proteins and other functions coded by the genome is a laborious process, and therefore the amount of characterized parts does not compare to the amount of genome sequences available. In the case of microalgal biotechnology, specifically synthetic biology, one area that we need to know more about is the functions of all proteins in a genome and how they interact as a system.

3. Protein functions – why we need them and how we get them

3.1 Bioinformatics and protein functions

For computational models to be useful for synthetic biology it is critical to have an extensive knowledge on protein functions. As can be seen by the low amount of wet-lab characterized proteins in online databases [55], this is still a laborious and in many cases difficult task. Therefore, bioinformatics is often used to predict protein function [56]. These predicted protein functions are directly and indirectly used in systems and synthetic biology, both for the insertion and deletion of genes [8, 10], and also for the creation of mathematical models used in systems biology research [57].

3.2 Retrieving protein functions through sequence similarity

Historically, electronic inference of protein functions is accomplished by looking for sequence similarity with proteins that have a known function. This has been practiced since 1985 with the FASTA alignment [58], the first algorithm that facilitated high-throughput protein sequence similarity searches. Similarly, the BLAST sequence alignment tool was published in 1990 and has been a staple tool in bioinformatics ever since [59, 60]. Most proteins annotated using computational methods have been done so using BLAST in one way or another. Manually curated annotations are often aided by BLAST sequence similarity searches, as are annotations predicted by high-throughput methods. Another tool that is ingrained in the world of genomics and protein

function is InterProScan [61]. This tools main functionality is the identification of protein domains given a protein sequence and a database of Hidden Markov Models [62], and meta-data attached to these domains in the InterPro database to annotate functions to these proteins using Gene Ontology (GO) terms [63-65]. The big downside of prediction methods based on sequence similarity, however, is the need of homologous proteins with reliable functional information.

3.3 Limitations of sequence similarity-based protein function prediction

Sequence similarity approaches rely on annotated proteins that have a close phylogenetic distance to the protein of interest. For example, the GreenCut2 is a resource of conserved proteins across photosynthetic plants and microalgae [66]. If a microalga has a protein that is present in the GreenCut2, and there is functional information on this protein available in another plant or microalgae, sequence similarity-based approaches will assume that the function can be transferred between these proteins. In the case of such conserved protein sequences, this is a reliable way of annotating protein functions. However, if a microalga has a protein that has no conserved sequence with a protein of known functionality, sequence similarity-based approaches will not be able to annotate a function to this protein. As discussed earlier, there is little experimental information available on microalgal proteins. Therefore, when using sequence similarity-based approaches to annotate their proteins, it will mostly return functions only for proteins that show sequence conservation with well-studied plant species such as *Arabidopsis thaliana*, as this is the organism that is the most well-annotated species in the plant and microalgal lineages according to SwissProt [55]. The result is a well-annotated core metabolism of a microalgae, but limited annotations for proteins that are unique for microalgae. Therefore, it is needed to additionally look at other approaches for annotating a function to microalgal proteins.

3.4 Multi-feature-based protein function prediction

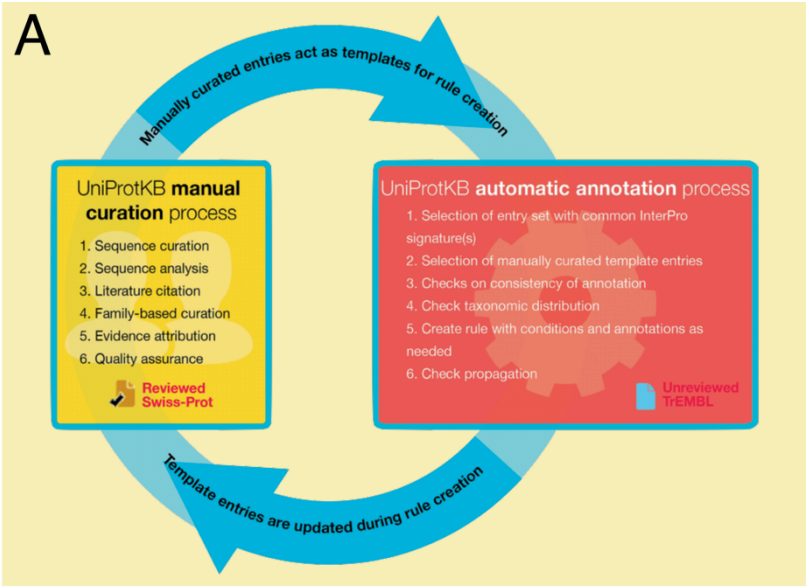
Another way of predicting protein functions is by training machine learning algorithms on a range of protein features [67]. Many machine learning techniques are able to correlate protein features such as amino acid composition, secondary structures, and disordered regions to GO terms, given a set of proteins with known functions and features [68]. A machine learning algorithm builds a statistical model around these features, which gives a likelihood that a certain pattern of features is linked to a certain protein function. In potential, these techniques are much more flexible than the earlier discussed sequence similarity approaches, as in principle they don't require sequentially closely related proteins to have a GO term annotation. These feature-based approaches try to understand protein function on a more fundamental level than transferring GO terms based on sequence similarity. Because of this they are complementary to sequence similarity-based methods.

3.5 Protein function information hubs

Currently, the best source for protein function information is SwissProt, part of the UniProt Knowledgebase [55]. SwissProt is a database of manually curated proteins and their functions, including meta data such as protein domains, links to other databases, and evidence codes for all the data associated to the protein. Alternatively, there is the TrEMBL database [55]. This database is also part of the UniProt Knowledgebase, but only contains computer-generated gene translations and protein functional predictions. How the data for these databases is derived is described in Figure 2A, and how this data can be used is described in Figure 2B.

The information in the UniProt Knowledgebase database is extremely useful for bioinformatics and systems biology research. Figure 2B describes how bioinformatics, systems biology, and synthetic biology research are all tied together, and data on protein functions plays a crucial role. This interaction is further discussed in **chapter 2 box 3**.

Additionally, there exist a number of databases that annotate proteins and their functional data to biochemical reactions and networks. These data are useful for the mathematical modeling of biological systems used in systems biology research, but also for the visualization and understanding of an organism its biochemistry. Databases such as KEGG [69], Reactome [70], WikiPathways [71], and MetaCyc [72] are invaluable to biotechnological research, and all of them require data on protein functions.


A
Manually curated entries act as templates for rule creation
UniProtKB manual curation process
1. Sequence curation
2. Sequence analysis
3. Literature citation
4. Family-based curation
5. Evidence attribution
6. Quality assurance
Reviewed Swiss-Prot
UniProtKB automatic annotation process
1. Selection of entry set with common InterPro signature(s)
2. Selection of manually curated template entries
3. Checks on consistency of annotation
4. Check taxonomic distribution
5. Create rule with conditions and annotations as needed
6. Check propagation
Unreviewed TrEMBL
Template entries are updated during rule creation

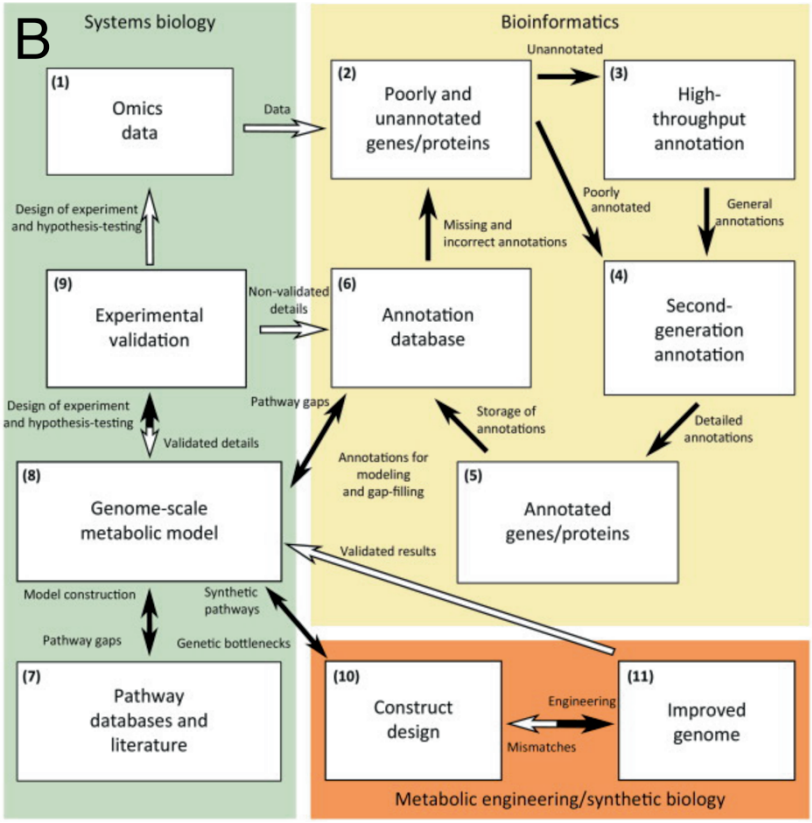

B
Systems biology
Bioinformatics
(1) Omics data
Data
(2) Poorly and unannotated genes/proteins
Unannotated
(3) High-throughput annotation
Poorly annotated
General annotations
Design of experiment and hypothesis-testing
Missing and incorrect annotations
(9) Experimental validation
Non-validated details
(6) Annotation database
(4) Second-generation annotation
Design of experiment and hypothesis-testing
Validated details
Pathway gaps
Storage of annotations
Detailed annotations
Annotations for modeling and gap-filling
(8) Genome-scale metabolic model
(5) Annotated genes/proteins
Validated results
Model construction
Synthetic pathways
Pathway gaps
Genetic bottlenecks
(7) Pathway databases and literature
(10) Construct design
Engineering
Mismatches
(11) Improved genome
Metabolic engineering/synthetic biology

Figure 2: A) An overview of how the SwissProt and TrEMBL databases of the UniProt KnowledgeBase retrieve their protein information. SwissProt proteins have been annotated using a well-defined manual curation pipeline to ensure that all manually annotated entries are handled in a consistent manner. Curation is performed by expert biologists using a range of tools that have been iteratively developed in close collaboration with curators. TrEMBL proteins are automatically annotated and not reviewed. B) An overview of the interaction between bioinformatics, systems biology, and synthetic biology research. Black arrows indicate predictions and literature data, and white arrows indicate experimental data.

4. Microalgae and protein functions

4.1 Limited knowledge on microalgal protein functions

As discussed earlier, to fully utilize microalgae as cell factories we would need to identify interesting metabolic proteins and circumvent metabolic constraints, which can be efficiently done using state-of-the-art genome editing techniques such as CRISPR-CAS9, and by creating new genetic circuits using synthetic biology. This requires a good understanding of the proteins involved and how they interact in a whole-cell system. However, there are only 267 experimentally characterized microalgal proteins in SwissProt as of 26-07-2018. Due to the inability to annotate protein functions due to a lack of homologous proteins with a characterized function, and because feature-based machine learning approaches are not as sequence similarity-based approaches, it is a challenge to generate a good annotation for microalgae on a genome scale. Therefore, we need a comprehensive utilization of bioinformatics and systems biology methodologies to understand the metabolic capabilities of microalgae. In particular, we need a good understanding of microalgal enzymatic functions in order to generate hypotheses as to how to circumvent metabolic constraints.

5. Botryococcus braunii

5.1 Why study this microalga

An interesting microalga to study is *Botryococcus braunii*. This alga is able to produce polysaccharides up to 50%, or hydrocarbons up to 40% of its cell dry weight depending on the strain. Additionally, it excretes the majority of these polysaccharides and hydrocarbons, making the harvesting of these products relatively easy. However, the wild-type of this microalga is a slow grower and lives in a community with various bacteria. This makes its commercialization capabilities limited. Using genomics research, we would be able to understand the metabolism of *B. braunii*, allowing us to potentially identify how this microalga makes a large number of polysaccharides and hydrocarbons, how it excretes them, and why its growth is limited. Understanding the features of this microalgae would allow us to genetically engineer the parts of the genome to faster growing and easier to handle organisms and identifying metabolic constraints would potentially allow us to circumvent the slow growth of *B. braunii*.

5.2 Studying *Botryococcus braunii:* Sustainable Polymers for Algae (SPLASH)

Because of the potential of *Botryococcus braunii* for biotechnological applications, an EU consortium was funded to study this microalga. The aim of this 4.5-year EU-project was to develop a microalgal cell factory able to sustainably produce hydrocarbons and exopolysaccharides, using *Botryococcus braunii* features and *Chlamydomonas reinhardtii* as a host organism. This project encompasses genomics and systems biology research to understand the product formation of two *Botryococcus braunii* races which make hydrocarbons or polysaccharides, the development of *in situ* extract and isolation methods for these hydrocarbons and polysaccharides, their conversion to commercial products, and a proof of concept of *Botryococcus braunii* cultivation on a pilot scale. Finally, based on information gained during this research, sustainability assessment and market analysis was done to assess the viability of *Botryococcus braunii* as a microalgal cell factory.

In this thesis research is aimed with *Botryococcus braunii* genomics in mind. In SPLASH, genomics and systems biology were used to understand the production of hydrocarbon and polysaccharide production in *Botryococcus braunii* on a fundamental level, with the goal of providing leads for cultivation

concepts, improved growth, and enhancement of hydrocarbon and polysaccharide production. To reach the objectives two *B. braunii* strains with either a strong affinity towards hydrocarbon or polysaccharide production, AC761 and CCALA778 respectively, were to be studied using comparative genomics. A comparative genomics study like this requires a thorough understand of the functions of metabolic proteins and transporter proteins. As discussed earlier, this is particularly challenging for microalgae. Therefore, many efforts were made to annotate functions to *Botryococcus braunii* proteins, and once we got those, to view these protein functions in the light of biological pathways and gene expression analysis coupled to the production of hydrocarbons and polysaccharides.

Thesis outline

The goal of this thesis is to contribute to increase our understanding of oleaginous microalgae as cell factories through developing and using bioinformatics tools and pipelines. Step by step I arrive at the point where I use bioinformatics and systems biology techniques to increase our knowledge of the metabolism of *Botryococcus braunii*.

In **chapter 2**, we provide an overview of the state of microalgae as cell factories. As we point out, during the time of publication for this review, only a small amount of existing genome sequencing was done for microalgae, and even less research on their biochemistry. In the meantime, many more genomes have become available, but the research on their biochemistry is a laborious process that does not increase as fast as current generation genome sequences. This means there is a low amount of experimental data on protein functions, and although there are a number of microalgal genome-scale metabolic models, these are based on homology with enzymes underlying core metabolic reactions of *Arabidopsis thaliana*. Even the microalgal model species *Chlamydomonas reinhardtii*, whose biochemistry is often used as a basis for other microalgal research, has a limited amount of experimental data on protein functions. In this chapter we suggest how to move forward with bioinformatics and systems biology to improve microalgae as cell factories.

In **chapter 3** we focus on a specific aspect of microalgal bioinformatics discussed in chapter 2: retrieving protein functions. We show that only about half of the microalgal proteins have an annotated function, and most functions annotated to microalgal proteins are transferred from *Arabidopsis thaliana* based on sequence similarity. Further, we review several protein function prediction methodologies and their pros and cons.

In **chapter 4** we present CrowdGO, a protein function prediction tool based on the wisdom-of-the-crowd principle. CrowdGO combines protein function predictions of three or more existing methods. We show that it performs significantly more accurate in GO term predictions than each method by itself, has a net positive effect in correcting existing predictions to its true prediction, and is able to annotate more proteins than each individual method.

In **chapter 5** we apply the method developed in chapter 4 to a real biological example. For this we use the oleaginous yeast *Cutaneotrichosporon curvatus*. Due to its distance to the model species *Saccharomyces cerevisiae* and its oleaginous nature, this species has two characteristics similar to microalgae:

distant phylogeny with well-characterized model species, and the ability to produce a lot of lipids. However, nearby species are better characterized than microalgae, resulting in a good middle-of-the-road species to test the tool in a real case scenario. With CrowdGO we get a stricter annotation of GO terms for the protein set. These annotations are used to get a general overview of yeast and its metabolism, and to use as a starting point in a biocuration pipeline for over a thousand metabolic and transporter reactions of *C. curvatus*.

In **chapter 6** we use our method developed in chapter 4 to perform a comparative genomics study between a mostly-lipid producing *Botryococcus braunii* and a mostly-polysaccharide producing *Botryococcus braunii*. We use quantitative proteomics and functional annotations, both from CrowdGO and manually annotated, to characterize several key pathways. Using this approach, we found several key enzymes which are related to the difference in lipid production and polysaccharide production between the two *Botryococcus braunii* strains.

Finally, **chapter 7** provides a general discussion on the findings of this PhD thesis.

Chapter 2

Green genes: bioinformatics and systems biology innovations drive algal biotechnology

Maarten J.M.F. Reijnders[1], Ruben G.A. van Heck[1], Carolyn M.C. Lam[1,2], Mark A. Scaife[3], Vitor A.P. Martins dos Santos[1,2], Alison G. Smith[3], and Peter J. Schaap[1]

[1] Laboratory of Systems and Synthetic Biology, Wageningen University, Dreijenplein 10, Building number 316, 6703 HB Wageningen, The Netherlands

[2] LifeGlimmer GmbH, Markelstrasse 38, 12163 Berlin, Germany

[3] Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK

Adapted from publication:

Trends in Biotechnology, December 2014, Vol. 32, No. 12

Abstract

Many species of microalgae produce hydrocarbons, polysaccharides, and other valuable products in significant amounts. However, large-scale production of algal products is not yet competitive against non-renewable alternatives from fossil fuel. Metabolic engineering approaches will help to improve productivity, but the exact metabolic pathways and the identities of the majority of the genes involved remain unknown. Recent advances in bioinformatics and systems-biology modeling coupled with increasing numbers of algal genome-sequencing projects are providing the means to address this. A multidisciplinary integration of methods will provide synergy for a systems-level understanding of microalgae, and thereby accelerate the improvement of industrially valuable strains. In this review we highlight recent advances and challenges to microalgal research and discuss future potential.

1. Diversity of microalgae and their biotechnological potential

Microalgae are simple photosynthetic eukaryotes that are among the most diverse of all organisms. Microalgae inhabit all aquatic ecosystems, from oceans, lakes, and rivers to even snow and glaciers, as well as terrestrial systems including rocks and other hard surfaces. Microalgae exhibit significant variation in physiology and metabolism, a reflection of the high level of genetic diversity that exists between different phyla owing to multiple endosymbiotic events, horizontal gene transfer, and subsequent evolutionary processes, producing a polyphyletic collection of organisms [54, 73]. Given this diversity, mining the genomes of these organisms provides a great opportunity to identify novel pathways of biotechnological importance. In particular, microalgae are of considerable interest for the synthesis of a range of industrially useful products, such as hydrocarbons and polysaccharides [74, 75], owing to rapid growth rates, amenability to large-scale fermentation, and the potential for sustainable process development [76].

Algae as a source of biofuel molecules, such as triacylglycerides (TAGs), the precursor for biodiesel [77], have been a focus in recent years, with potential yields an order of magnitude greater than competing agricultural processes [78]. Evaluations of current technologies demonstrate that microalgae are commercially feasible for biofuel production, but are not yet cost-competitive with petroleum products [79, 80], the metric upon which commercial success ultimately lies. For example, the net energy input versus output for large-scale algae biodiesel production was estimated to be 1.37, compared to 0.18 for conventional/low-sulfur diesel [79]. Currently, for microalgae to synthesize TAG it is necessary to expose them to stress conditions such as nutrient limitation, which reduces growth and increases energy dissipation. The trade-off between biosynthesis of TAG and cell growth is therefore a severely limiting factor [81]. If a better understanding of the metabolic and regulatory networks were available, they could be rewired for increased TAG synthesis, with fewer drawbacks than for existing algal cells.

The production of other interesting algal products will also benefit from a better understanding of microalgae at a systems level. For example, polysaccharides such as starch and cell wall materials can be used for biotechnological applications [82]. These carbohydrates can be degraded to fermentable sugars for bioethanol production [83], or serve as chemical building blocks for renewable materials, but the composition and proportions of the different sugar components require optimization. Similarly, various

valuable secondary metabolites produced by microalgae are of interest in the food, nutrition, and cosmetics industries [74], but often they are produced in trace amounts, or only under conditions that are not amenable to industrial cultivation.

Over 30 microalgal genomes have been sequenced, and numerous transcriptomics, proteomics, and other systems-biology studies have been performed. Nevertheless, our understanding of metabolic pathways within these microalgae remains limited [84]. Significant knowledge gaps need to be filled between omics data, the annotation thereof, and our systems-level understanding. This will allow the conversion of these resources into usable genome-scale models (GSMM) and provide the basis for effective metabolic engineering, synthetic biology and biotechnology. We consider here the potential application of advanced methods to improve the functional annotation of algal omics data, to increase the resolution of GSMM, and ways to integrate available computational methods for effective exploitation of microalgae in biotechnology.

2. Annotation challenges for microalgae

The nuclear genome of the green alga *Chlamydomonas reinhardtii*, sequenced in 2007 [54], is approximately 120 Mb and comprises some 15 000 genes. Although *C. reinhardtii* is commonly used as a reference for the annotation of other microalgae, only a subset of ~50 proteins have an experimentally validated function according to the UniProt database (http://www.uniprot.org), compared to 6800 proteins for the model plant *Arabidopsis thaliana*. Consequently, most *C. reinhardtii* genes have been computationally annotated by inferred homology with *A. thaliana*, and other plant species and microbes [54], using BLAST (basic local alignment search tool) or family-wise alignment methods such as HMMER and InterProScan (Table 1). BLAST-based methods often use the principle of one-to-one recognition, meaning that annotation of a query gene is based on the annotation of a single known gene. This limits the success rate for recognition and correct functional annotation of the more distantly related *C. reinhardtii* genes but becomes even more problematic when the *in silico*-derived functional annotation of *C. reinhardtii* is subsequently used for annotation of other algal species. This is because, owing to a lack of common ancestry, two algal species can be more diverse than, for example, any two plant species. Therefore, these methods, which are highly suitable for high-throughput analysis because of their simplicity, are less appropriate for

accurate in-depth annotation of algal genomes. In the CAFA (critical assessment of protein function annotation) experiment [56], the accuracy of more advanced functional annotation algorithms was assessed. The CAFA concluded that 33 of 54 tested functional annotation algorithms outperformed the standard BLAST-based method (Table 1). The substantial improvement can be explained by the fact that these second-generation methods do not apply the one-to-one recognition principle but, to increase their success rate, use instead a one-to-many recognition strategy and/or include context-aware principles for annotation. An example is Argot2 (Box 1) [2], which applies the one-to-many recognition strategy by calculating the statistical significance of all candidate homologous genes found by BLAST [85] and HMMER [86], combined with an assessment of semantic similarities of associated GO terms. In a context-aware multilevel approach, annotation is not merely based on sequence similarity, but other factors such as protein–protein interactions [87], transcript expression patterns [87], phylogenetic trees [88], compartmentalization information [89], and literature [90] are also taken into account. FFPred2 from UCL–Jones [91] is the prime example of such a homology-independent functional annotation algorithm.

Table 1: Features of commonly used functional annotation tools

Methods	Success rate[a]	Computational speed	Availability	Additional notes	Refs
Standard BLAST	Limited	Fast	Online/offline	Dependent on global sequence similarity rate for success Suitable for high-throughput analysis	[85]
HMMER	Moderate	Fast	Online/offline	Family-wise alignment method Suitable for high-throughput analysis	[86]
InterProScan	Moderate	Slow	Online/offline	Family-wise alignment method Uses pre-computed protein domains	[61]
FFPred2	High	Slow	Limited online/offline	Algorithms currently trained on non-algal datasets Not suitable for high-throughput analysis	[89, 91]
Argot2	High	Moderate	Limited online	Initial selection is dependent on BLAST and HMMER output Additionally predicts compartmentalization User-friendly interface	[2]

[a] For distantly related sequences

Box 1. Argot2

One of the top performers in the CAFA experiment is Argot2 (annotation retrieval of gene ontology terms) [2]. It stands out in terms of simplicity, as well as by incorporation of BLAST and HMMER. Argot2 combines an easy interface with multilayer analysis, making it a perfect starting point for biologists wishing to annotate their data.

Argot2 requires a nucleotide or protein sequence as input. It queries the UniProt and Pfam databases using BLAST and HMMER respectively, providing an initial high-throughput sequence analysis. A weighting scheme and clustering algorithm are then applied to the results to select the most accurate gene ontology (GO) terms for each query sequence. The user can choose to perform this entire process online at the Argot2 webserver, limited to one hundred sequences per query. Alternatively, if the BLAST and HMMER steps are performed locally and provided to the webserver, over 1000 sequences can be submitted per query. After the analysis is completed, which can take several hours depending on the amount of input data, the user is provided with the prediction results as well as the intermediate BLAST and HMMER files. These predictions include molecular function, biological processes, and cellular component GO terms for each query. Predicted GO terms are ranked by a score based on statistical significance and specificity. Optionally, the user can choose to compute protein clusters based on functional similarity.

Advanced multilevel annotation methods effectively increase the recall of function prediction while maintaining an acceptable precision. The challenge in genomic annotation for microalgae lies in the small number of experimentally validated algal genes and the lack of algae-specific contextual data such as protein interaction and compartmentalization data. This results in a relatively low number of genes that are predicted to have a specific biological function. To overcome this, multiple annotation methods and data sources should be combined. The combined result increases the number of annotated genes, while a consensus prediction among the different methods improves the accuracy of the annotation [92]. Owing to their simplicity and speed, first-generation methods can be used for initial high-throughput analysis of a large set of genes. Second-generation methods can then be used for a refined analysis of these genes. However, to utilize these advanced methods fully, a significant amount of experimentally determined contextual data is required. Although increasing amounts of gene expression data are being generated, little structural and protein interaction data are being generated for algae. In the absence of such experimental facts it is still possible to generate this contextual information by *in silico* prediction methods [91, 93], but whilst studies have shown that this is a feasible option [94], caution is necessary because there is a high risk of error propagation.

Apart from functional annotation it is also important to establish the cellular location of a protein. For this there are several tools available, including Argot2 (Box 1) [2], TargetP [95], SignalP [96], PSORTb [97], and PredAlgo [98]. The last is a tailor-made multi-subcellular localization prediction tool dedicated to three compartments of green algae: the mitochondrion, the chloroplast, and the secretory pathway. However, owing to the limited number of algal proteins with a known cellular localization, which can be seen for example from the quantitative subcellular localization of roughly 80 proteins [99], or the collection of roughly 1000 chloroplast-localized proteins from *C. reinhardtii* [100], the algorithm is trained with a relatively small *C. reinhardtii* dataset [98]. This raises questions regarding reliability for other algal species because the polyphyletic nature of different microalgae means some algal species are distantly or not related, and this can result in a different subcellular localization of homologs. Therefore, it is advisable to use PredAlgo in combination with non-algal-specific tools in a similar way as for functional annotation.

To support large-scale annotation of algal sequence data, up-to-date databases and readily available supporting tools are required. Online databases provide the means to share data easily such that the scientific community can profit as a whole. Supporting tools can assist in annotating genes, pathways, and performing statistical analysis. While genomic data for various algae are available in NCBI and UniProt, the amount of public data is lagging behind in comparison to plant and bacterial species. In addition, tools and databases that do more than storing the available sequencing data are needed. A small number of tools are available, although these are often limited to *C. reinhardtii*. One such tool is ChlamyCyc [101], a *C. reinhardtii*-specific pathway/genome database of the MetaCyc [72] facility for metabolic pathway analysis. A peptide database, ProMEX, is available that contains over 2000 *C. reinhardtii* peptides which are usable for proteomics analysis [102]. In addition, the Augustus tool, which is commonly used for prediction of eukaryotic genes [103], has a tailor-made section for *C. reinhardtii*. Finally, the Algal Functional Annotation Tool [104] incorporates annotation data for a few microalgal species from several pathway databases, ontologies, and protein families. Broadening the scope of these annotation tools for a range of microalgae would allow comparative analysis, which is useful for easy mapping of various differences between microalgae. In this context, a useful tool which has been applied to plant research is Phytozome (http://www.phytozome.net) [105], a comparative hub for analysis of plant genomes and gene families. It acts as a reference for the key data of many plant species and provides click-to-go features such as BLAST and summaries key data. Phytozome has grown to be a major asset to the plant science community. Although it contains data from a few green algae, an expanded web-portal focused on algal systems-bioinformatics research could be of immense benefit to the field, particularly for those studying the more industrially relevant diatoms and heterokont species (Table 2). Such a web-portal would provide access to new and existing tools specifically useful for algal species and facilitate exposure to a broad audience. In addition, it could act as a hosting platform for small but useful tools such as a refined algal literature research algorithm and tools that suggest genes to fill gaps in metabolic or regulatory pathways for microalgae. Adopting an algal web-portal would provide a good overview of all available data and tools and help to reduce the redundancy that is often seen in biology and bioinformatics.

Table 2: A list of selected industrially useful microalgae

Species	Genome size (mb)	Available proteins	Reported industrially relevant characteristics	Refs
Chlamydomonas reinhardtii	120	15,144	Model system for unicellular green algae	
Monoraphidium neglectum	68	16,761	Up to 21% dry weight neutral lipid under nitrogen starvation	[106]
Nannochloropsis gaditana	34	15,361	Can produce high amounts of omega-3 long-chain polyunsaturated fatty acids Up to 50% dry weight oil content	[75] [107]
Nannochloropsis oceanica	28	242		
Phaeodactylum tricornotum	27	10,673	Can produce antibacterial fatty acids (9Z)-hexadecenoic acid (palmitoleic acid; C16:1 n-7) and (6Z, 9Z, 12Z)-hexadecatrienoic acid (HTA; C16:3 n-4)	[107]
Chlorella variabilis	46	9,831	The first sequence Chlorella genome A model genome for understanding other chlorella species	[108]
Ostreococcus tauri	12.6	9,050	Smallest sequenced microalgal genome with simple cellular structure	
Chlorella protothecoides	22.9	7,039	Up to 55% dry weight lipid content in heterotrophic growth Highest published biomass yield, average 3.37 g/dw $L^{-1} h^{-1}$ in heterotrophic growth	[75, 109, 110]
Chlorella vulgaris	N.a.	292	Up to 42% lipid content in photobioreactor with artificial waste water Up to 26% total lipid in dry weight in heterotrophic growth	[108, 111]
Dunaliella salina	N.a.	238	Up to 10% carotenoids in dry weight 90% beta-carotene in carotenoids	[112]
Haematococcus pluvialis	N.a.	60	Highest reported yield of antioxidant astaxanthin (3.8% dry weight)	[113]
Botryococcus braunii	~166-211	30	Up to 57% total lipids in dry weight Contains exopolysaccharides	[114-116]
Neochloris oleabundans	N.a.	0	Up to 56% total fatty acids in dry weight under nitrogen-deprivation	[81]

3. Understanding algal metabolism at a systems level

The sheer number of genes for metabolic enzymes, combined with the complexity of cellular metabolism, means that it is not straightforward to establish metabolic capability, even for well-annotated species. This limitation has led to the development of metabolic models which represent a snapshot of metabolism of an organism in a network format. Once an annotated algal genome or transcriptome is available, a corresponding genome-scale metabolic model (GSMM) can be reconstructed and the topology of the metabolic network of the algal species can be analysed. An initial draft model can be generated directly from the genomic annotation and is then adjusted and expanded based on experimental data, literature, and gap-filling procedures. The final model then includes all reactions the alga is known to perform as well as the associated genes and constraints, for example, reaction directionalities and rate limits. Owing to their comprehensive representation of metabolism, metabolic models form the basis for a large and diverse set of mathematical methods for predicting metabolic behaviour. These methods include the widely employed flux balance analysis (FBA) [117] and flux variability analysis (FVA) [118], but also methods integrating fluxomic, transcriptomic, or proteomic data (Box 2) [119]. For an extensive overview of mathematical methods using metabolic models we refer to Zomorrodi *et al.*[120]. We focus here on recent developments in the modeling of microalgae specifically.

Metabolic models of microalgae reflect the modeling counterpart of their current annotation; therefore, inconsistencies between model predictions and experimental findings indicate missing and/or poor annotations. For example, experimentally identified metabolites were compared to metabolites that could be produced in metabolic reconstructions of *C. reinhardtii* [121, 122] (Table 3). Metabolites found experimentally but not in the models initiated pathway elucidation and identification of the corresponding genes, and thereby led to an improved genomic annotation [121]. This procedure was automated by Christian *et al.* who designed a gap-filling method to identify reactions allowing production in a model of experimentally detected metabolites [122]. These updated reactions and annotations [121, 122] were subsequently stored in ChlamyCyc [101], allowing continuous expansion of the database. Concurrently, a separate *C. reinhardtii* metabolic model, iAM303, was created in which the included open reading frames were experimentally validated. This led both to improved structural genomic annotation and to additional support for the reactions included in the model [123]. This model was greatly expanded in iRC1080 in 2011 and additional

ORFs were validated [124]. The predictive power of the latter model was tested for 30 environmental conditions and 14 gene knockouts. In addition, iRC1080 predicted essential genes (lethal phenotype upon knockout) under different experimental conditions, although these predictions remain to be validated [124]. Recently GSMMs for *Ostreococcus tauri* and *Ostreococcus lucimarinus* have been constructed [125] (Table 3), demonstrating expansion in the field. The initial models, based on the available gene annotations, revealed that these could not account for the production of many biomass constituents [125]. The gap-filling method designed in [122] was subsequently employed to find suitable reactions for the production of these metabolites [125].

It is well recognized that the exact choice of growth conditions is highly important in attaining desired metabolic activities. Metabolic models can explore how different growth conditions affect metabolism and can identify theoretically optimal conditions for a given metabolic objective. For example, multiple metabolic models of *C. reinhardtii* were used to simulate metabolism under autotrophic, heterotrophic, and mixotrophic conditions to verify model predictions [46], to investigate how metabolite production is influenced 46, [119], and to contrast mutant strains [124]. *C. reinhardtii* metabolic models were also used to determine how the quantity of light [124, 126, 127] and its spectral composition [124] affect metabolism. Of particular interest is the possibility to predict an optimal light spectrum for a given metabolic goal [124]. In contrast to these successful models of *C. reinhardtii*, the metabolism of other algae is only poorly understood. For example, some industrially relevant algae can currently not be grown efficiently without bacterial presence [128]. Potentially, these algae and associated bacteria can be modeled simultaneously to deduce their relationship, as has been done for other microbial communities [129, 130].
The most comprehensive algal metabolic models to date are iRC1080 [124] and AlgaGEM [46], which are GSMMs and account for various cellular compartments. However, they vary in degree of compartmentalization (Table 3). In iRC1080, half (865/1730) of the non-transport reactions occur in cellular compartments other than the cytosol. By contrast, this is only about 12% (201/1617) for AlgaGEM. This reflects the fact that independently generated GSMMs for the same organism can differ significantly in their representation of metabolism because different sources of information are included. By combining the information from all currently available *C. reinhardtii* metabolic models, as well as from improved annotation methods, a single and more-comprehensive GSMM may be obtained. This consensus *C. reinhardtii* GSMM would be an important starting point for the

generation of GSMMs for other interesting microalgae, with the proviso mentioned earlier that it might not be applicable to distantly related microalgae. Alternatively, *ab initio* models can be made using genomic data for the alga in question, but employing the strategies and tools developed for *C. reinhardtii*, as has been done for *Ostreococcus*[45]. Ultimately, GSMMs of various microalgae will be valuable for designing strategies that increase the production of compounds of interest [120, 131]. This, combined with the design of novel synthetic pathways, such as the species-independent prediction demonstrated for novel isobutanol, 3-hydroxypropionate, and butyryl-CoA biosynthesis [132], will pave the way for model-driven engineering of algal species

Table 3: Overview of metabolic models of microalgaea

		Model									
		Christian et al [122]	Boyle and Morgan [119]	iAM303 [123]	AraGEM [46]	Alga GEM [46]	Cogne et al [127]	iRC1080 [124]	Kliphuis et al [126]	Krumholz et al [125]	
Year		2009	2009	2009	2010	2011	2011	2011	2012	2012	
Species		C. reinhardtii (C. r.)	C. r.	C. r.	A. thaliana	C. r.	C. r.	C. r.	C. r.	O. tauri	O. lucimarinus
Parent model (if any)		-	-	-	-	Ara GEM	Boyle and Morgan	iAM303	Boyle and Morgan	-	-
Total number of reactions		~1,200	788	279	1,601	1,718	314	2,191	159	871	964
Distribution (%) of biochemical reactions among major compartment-spanning transport reactions	Cytosol	N.a.[c]	21.57	17.67	84.14	85.26	95.39	40.82	84.08	99.30	99.37
	Chloroplast		34.52	23.69	5.84	5.84	0	22.75	3.82	99.30	99.37
	Mitochondria		11.29	14.06	1.97	1.95	0	10.29	0	0	0
	Peroxisome/ Glyoxysome		0	4.42	2.66	2.36	0	2.17	0	0	0
	Others[b]		0.13	4.42	0.06	0.06	0.33	4.77	1.27	0.12	0.10
	Transport		32.49	35.74	5.33	4.89	4.28	19.21	10.83	0.58	0.52
Total number of genes from model species		[d]	243	174	1,403	843	[d]	1,086	41	[d]	[d]
Number of unique decompartmentalized metabolites		~1,500	261	113	1,509	1,645	277	1,071	138	1,009	1,105
Number of distinguished biological cellular compartments		0	3	6	5	5	1	9	2	1	1

[a]The total number of reactions, total number of genes, unique decompartmentalized metabolites, and biological cellular compartments were taken from available model files and/or supplementary documents of the corresponding references. The distributions of biochemical reactions among different compartments as well as compartment-spanning transport reactions are shown as the percentage of their sum.

[b]The category 'others' refers to the following compartments: flagellum, Golgi apparatus, thylakoid lumen, nucleus, and eyespot. [c]N.a., not available.

[d]Gene information not available from model files nor supplementary documents.

Box 2: Flux analysis in microalgae

Flux balance analysis (FBA) [117] is the most commonly applied method to simulate metabolism in genome-scale metabolic models. It identifies as a theoretically optimal use of metabolic capabilities for a selected metabolic objective in a specific environment. Because some microalgae can grow autotrophically in chemically defined medium, the boundary conditions for consumption of all medium components are well specified in those cases. This is advantageous for *in silico* metabolic flux analysis using metabolic models to address, for example, how a microalga can achieve maximal growth under defined illumination. In addition, disabling the metabolic capabilities associated with a gene allows simulation of mutant strains. FBA can thus assess the potential of different strains and different environmental conditions. To run FBA, all reactions are organized in a stoichiometric matrix *S*. Each column in *S* represents a different reaction, and each row a different metabolite. A nonzero value at position [*i*, *j*] thus indicates the stoichiometric coefficient of metabolite *i* in reaction *j*. FBA then employs two different constraints. (i) Metabolism is assumed to be in steady-state; production/degradation of intermediate compounds is not possible, and (ii) thermodynamics (reversibility) and substrate availability both dictate lower and upper flux bounds for individual reactions. Finally, one or more reactions are selected to represent the metabolic objective, for example, algal biomass production. Together, the *S* matrix, the constraints, and the objective function form a linear programming problem:

max(**X***__c__)

s.t. *S*__x__ = ***o***

x >= **lb**

x <= **ub**

where ***x*** is the flux vector, ***c*** is the objective vector, **o** is a null vector ensuring steady-state, and ***lb***/***ub*** are the lower/upper bounds for each reaction. The vector ***x*** represents a flux distribution with the theoretically maximal value for the metabolic objective. However, because of the presence of alternative/cyclic pathways, there are often alternative flux distributions with equally high values for the objective function. Flux variability analysis [118] explores for each reaction to what extent the flux can vary while permitting only a small reduction in the obtained value. In addition, experimental data can be used to provide additional constraints. For example, ^{13}C-labeling experiments provide experimentally measured fluxes as inputs for the model simulations [133]. Several FBA-based methods also facilitate the integration of transcriptomic, proteomic, and metabolomics data with metabolic models to constrain reactions based on measured RNA or protein levels [86,87[134]. Thereby, flux distributions are identified which are most consistent with the expression data [135]. Because of the greater number of quantitative genome-wide transcriptomic studies compared to those analyzing the proteome,

applications using transcriptomic data have been relatively abundant. However, the methods generally do not distinguish between these two types of data, and metabolic models can therefore be integrated with, and their predictions compared to, experimental data yielding new insights into metabolic functioning.

4. Integrating bioinformatics and modeling for algal biotechnology

The GSMMs provide a basis for both computational and laboratory-driven experiments, assisting in the discovery of biotechnology-driven solutions for genetic bottlenecks in algae. For example, to enable microalgae to become a viable industrial biosynthesis platform, their photosynthetic efficiency, product yield, and their growth rates under conditions for product synthesis will need to be addressed. Photosynthetic efficiency, with an estimated maximum of 8–9% in wild type algae [42, 136], sets a limit to both product synthesis and growth rate. Because of efficient light-harvesting antenna, algal cells can absorb much more light than they are able to use for photosynthesis [136], with the excess being lost as heat or fluorescence. In dense algal cultures, such as might be found in industrial cultivation systems, this reduces light penetration, placing a limit on the depth of the culture, increasing the surface area to volume ratio required for maximum productivity. Truncated light-harvesting chlorophyll antenna size (*tla*) mutants of *C. reinhardtii* with reduced antenna size have been shown to have improved solar energy conversion efficiency and photosynthetic productivity in mass culture and bright light [137]. Another study has modeled different pathways for the process of carbon fixation [138] as a means to overcome the low oxygenase activity of Rubisco [139]. Bar-Even *et al.*[138] computationally identified alternative carbon fixation pathways by using approximately 5000 known metabolic enzymes, hoping to find carbon fixation pathways with superior kinetics, energy efficiency, and topology. Some of their proposed pathways were estimated to be up to two- to threefold more efficient than the conventional Calvin–Benson cycle. Using an algal GSMM to study these pathways would help in understanding how these predictions may affect biomass and product synthesis in microalgae.

As explained earlier, nitrogen limitation is a necessary stimulus for TAG accumulation by microalgae [81]. This also triggers a reduction in photosynthetic membrane lipids and cessation of cell growth. The link between accumulation of lipid (including TAG) and macronutrient stress has been investigated using a systems approach, such as in a proteomic analysis of *C. vulgaris*, which led to identification of new transcription factors associated with lipid accumulation, offering the prospect of TAG overproduction independently of nutrient limitation [140]. In another approach, in the diatom, *Thalassiosira pseudonana*, TAG production was increased not by targeting the biosynthesis of lipids, or the production of competing energy sinks, but instead by RNAi knockdown of lipases involved in glycerolipid catabolism [141]. The integration of knowledge gained from GSMMs and similar metabolic engineering offers scope for improved efficiency based on rational

design. For example, farnesyl pyrophosphate is a precursor of terpenoids, steroids, and carotenoids, and the metabolite itself is also a product of interest in algae. Bacterial promoters responsive to the toxic accumulation of farnesyl pyrophosphate have been identified and used to regulate the expression of the precursor biosynthesis operon. This increased the yield of amorphadiene twofold over chemically inducible and constitutive gene expression [142]. Such an approach in microalgae would be foreseeable in the future, when promoters in various algal species are better understood, through model-driven design that incorporates systems data.

Alongside genomic sequence information, a key requirement is the ability to carry out genetic transformation, and while this is routine for *C. reinhardtii*, and a few other species such as the diatom *P. tricornutum*, in the past few years there has been a rapid increase in published methods for the transformation of several species of industrial interest including *Nannochloropsis* sp. [143]. Moreover, the ability to engineer the chloroplast genome offers considerable opportunities for metabolic engineering, given the focus of this organelle on biosynthesis [144]. Nevertheless, for predictive metabolic engineering there is an urgent need to expand the toolbox, particularly for the regulation of transgene expression. In this context, there are several well-established systems for inducible gene expression in *C. reinhardtii*, most notably promoters that are regulated in response to nitrate (*NIT1* or *NIA1*) [145] or copper (*CYC6*) [146]. More recently, vitamin-responsive *cis* elements have been identified, namely a cobalamin (vitamin B_{12})-responsive promoter [147] as well as a thiamine (vitamin B_1)-responsive riboswitch [148], and these have been demonstrated to be useful regulatory tools. Vitamins have the advantages of being benign, cheap, and effective at low concentrations. However, the majority of these elements have been discovered by coincidence rather than by design, and a more rational approach will come from use of transcriptomic data to provide promoters responsive to particular regulators, for example in response to CO_2 levels [149]. Further facilitation of transgene expression comes from the use of 2A peptides [150] which cause self-cleavage to release individual domains from a fusion protein. They thus provide the capacity for operon-like transgene expression within the nucleus. Marker recycling methods for chloroplast engineering have also been developed for *C. reinhardtii* [144, 151]. However, despite these developments, progress remains parallel in nature and heavily focused upon the development of *C. reinhardtii*. Information from algal genomes will be key to increasing the molecular tools available.

Nonetheless, for microalgae to be developed as a commercially viable biotechnology platform, rational design to address the current shortcomings must be achieved through the development of fit-for-purpose metabolic engineering or synthetic-biology resources. The diversity of algae provides considerable biotechnological potential but also presents a serious challenge to establishing common tools and approaches. The relative immaturity of the field, combined with the enticing potential of integrating predictive design of microalgae with the bioinformatics and systems-biology modeling framework (Figure I in Box 3), offers new perspectives for future improvements in algal biotechnology. By adapting cutting-edge developments in functional annotation for microalgae, and using these for the modeling of their metabolic and regulatory pathways, it will be easier to establish common features of algal genomes, and at the same time identify novel pathways for exploitation. A more accurate and elaborate functional annotation of omics data by combining first- and second-generation methods will allow reverse-engineering based on algal genome-scale metabolic models. These can then be used to inform hypothesis-driven metabolic engineering experiments in microalgae. Such an integrated approach is currently missing, but will provide the knowledge necessary for predictive modifications of algal industrial biotechnology platforms in the future.

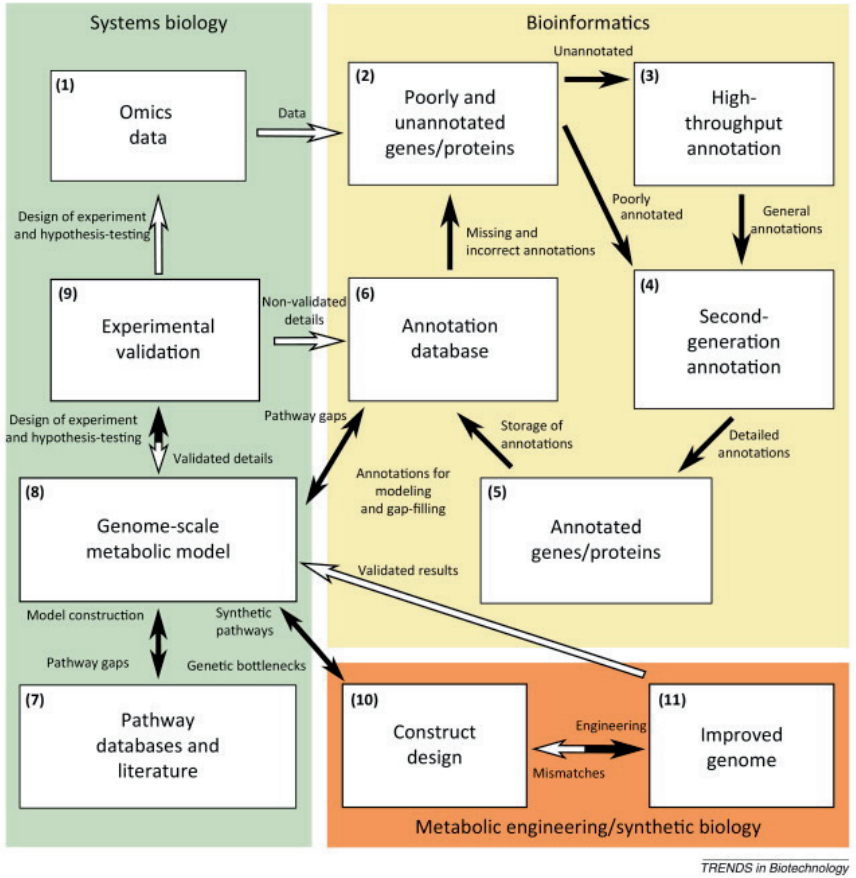


Figure 1: A multidisciplinary workflow integrating bioinformatics, systems biology, and metabolic engineering/synthetic biology of microalgae. Black arrow, *in silico* data or predictions; white arrow, experimental (wet-lab) data.

Box 3: Integrative and systematic understanding of algae

Improvements in algal annotations will need to interact closely with systems modeling of the metabolic and regulatory networks to refine our understanding of the capabilities of a specific alga and to provide a basis for applications in biotechnology. Figure I shows the connection between the various stages in bioinformatics and systems-biology modeling. New algal genomic, transcriptomic, and proteomic data are collected (step 1), allowing the identification of genes and proteins (step 2). After first-generation high-throughput functional annotation (step 3), a refinement step using secondgeneration functional annotation algorithms (step 4) is applied. The bioinformatics annotation itself is an iterative process for genes and proteins until they are deemed sufficient (step 5). These annotations (step 6), as well as data available from public databases and the literature (step 7), are then used by systems-biology modeling to reverse-engineer a GSMM (step 8) to study metabolic interactions in different circumstances in detail. After attaining a GSMM, experimental validation of the metabolic model (step 9) should be performed to validate model predictions or pinpoint inaccuracies and knowledge gaps. Depending on these results, additional omics data or refinements of annotation are required. Owing to the low number of experimentally validated algal proteins, the feedback loop from algal modeling back to genes/proteins function prediction plays a significant role in strengthening the knowledge foundation, and this will ultimately underpin efficient engineering of algal genomes for industrial product synthesis. Once an algal GSMM is constructed it should be made available in a common public database and literature.

Concluding remarks

The significant gap of unknown and non-validated gene and protein functions in algae remains one of the top challenges faced by scientists wanting to tap further into the potential of these organisms for sustainable biosynthesis. Predictive design of metabolic engineering strategies for microalgae still has a long journey ahead. An improved understanding of the metabolism, regulation, and growth of algae, together with their interactions with coexisting bacteria, is a crucial first step. Extending bioinformatics approaches for function prediction through incorporation of new methodology, integrated and flexible databases, in combination with metabolic modeling and model-driven design of experiments at the systems-biology level, will underpin this process and enable the future era of algal industrial biotechnology.

Acknowledgments

We acknowledge support from the European Commission 7th Framework Program (FP7) project SPLASH (Sustainable PoLymers from Algae Sugars and Hydrocarbons), grant agreement number 311956.

Chapter 3

Algal Omics: The Functional Annotation Challenge

Maarten J.M.F. Reijnders[§], Benoit M. Carreres[§] and Peter J. Schaap[*]

Laboratory of Systems and Synthetic Biology, Wageningen University, Wageningen, The Netherlands

*Address correspondence to this author at the Laboratory of Systems and Synthetic Biology, Wageningen University, Dreijenplein 10, 6703, HB, Wageningen, The Netherlands; Tel: +31 317 482105; E-mail: peter.schaap@wur.nl

[§]These authors contributed equally to this manuscript.

Adapted from publication:

Current Biotechnology, **2015**, Volume 4, No. 4

Abstract

Background: To fully exploit the potential of microalgae as commercial green hosts, the scientific community has to improve their understanding of these organisms from a systems biology perspective. Compared to other model organisms, our genomic knowledge of the microalgae model species *Chlamydomonas reinhardtii* is very limited. Currently, almost 90% of the functional annotated proteins of *C. reinhardtii* and of other microalgal proteins are homologs of *Arabidopsis thaliana* proteins, which suggests that for the most part only the metabolic core conserved between these species is properly annotated.

Objective: This review highlights how proteins outside of this core can be annotated by applying publicly available tools and methods. These include the use of novel state-of-the-art prediction tools, combinations of these tools, and the use of metabolic modeling-assisted functional annotation. Furthermore, we discuss the need for data on the subcellular location of microalgal proteins. Finally, some remaining bottlenecks regarding functional annotation of microalgal proteins are discussed.

Conclusion: We conclude that both large dry-lab and wet-lab efforts are required to generate reliable functional annotations of microalgae.

Keywords: Microalgae, bioinformatics, systems biology, annotation, genomics, proteomics, protein function.

1. INTRODUCTION

Microalgae are considered as promising organisms for a bio-based economy and unlocking their power potentially holds solutions for achieving global sustainability. In order to cope with some of the most demanding nutritional and energetic challenges of the future, research has focused on the renewable oil that can be extracted in significant amounts from these microalgae to create sustainable consumer products. However, compared to the more traditional sources, economically interesting molecules, such as triacylglycerides and polysaccharides, are currently not produced at a cost competitive rate [42]. To increase the yield, it is important to understand the genomic makeup of microalgae. More specifically, it is important to understand microalgae as biological systems at such a level of detail that mathematical models can be developed for these cell factories. These models can then predict the most optimal conditions for growth and production of interesting compounds and can guide genetic precision engineering of these cell factories [152]. Such models, often in the form of genome-scale metabolic models, require a thorough functional annotation of the proteins encoded by the genomes.

In today's age of biology, computational annotation of protein functions is of vital importance. Sample throughput of the classical biochemical and genetic methods is simply too low to be considered as an alternative. However, there is large phylogenetic distance between microalgae and well- characterised (model) species [153], and this distance hampers standard computational methods for genome annotation. Many of the popular computational methods for function prediction try to infer homology by calculating sequence- based statistical similarity scores with proteins of known function [61, 154]. This works fairly well for a comparison between a well-studied model organism with a large set of proteins validated by biochemical and genetic methods, such as *Arabidopsis thaliana* and *Escherichia coli*, and close by plants and bacteria, but the efficiency of a sequence similarity based annotation method decreases drastically when it is used between a group of species with little experimentally validated proteins, or when it is used for species that have a large phylogenetic distance to a well- studied homologous species. The most studied microalgae species *Chlamydomonas reinhardtii* became known early on as an excellent model species for microalgae because of its genetic amenability [155], but two decades later our genetic knowledge of this species still trails far behind that of other model species. Currently only some 150 proteins are characterized by direct biochemical methods. Furthermore, due to the large phylogenetic

distance to the closest well- studied model organism [153], *Arabidopsis thaliana*, only the most conserved genes are properly computationally annotated. Subsequently, only this limited set can be used as a reference set in sequence similarity-based methods to annotate other algal species of interest. As a result, only the conserved core metabolism of various microalgae is functionally annotated with a high level of confidence. Most of these microalgae were, however, selected for their ability to produce interesting and novel compounds [156]. To truly exploit microalgae for a bio-based economy, it is therefore important to know the function of the proteins that are not part of this metabolic core. By gaining more detailed genomic knowledge we will be able to produce more accurate algae specific genome-scale metabolic models. This allows for the prediction of biomass composition and conditions for optimal growth rates of microalgae, as well as for diversifying between the unique characteristics and capabilities of different microalgae and strains. Recently, alternative methods to functionally annotate microalgae have been described [157]. In this review we assess the current state of microalgal functional annotation, standardly used methods and discuss some alternative methods and workflow based on novel annotation tools that are currently available to the scientific community. Finally, we address some bottlenecks that currently cannot be solved by computational methods.

2. AVAILABLE DATA

From early on, *Chlamydomonas reinhardtii* was the only microalgae species that was extensively studied on a molecular scale. This species was first proposed a model organism for algal genetics in 2001 [155], and a draft genome sequence was available in 2003 [158]. However, due to "unusual challenges" in generating a high-quality genome [158], the genome was only published as late as 2007 [54]. For *C. reinhardtii* to serve as a model species to which other algae can be compared, it is important that many algal- specific protein functions and other key functions are based on experimental evidence, and not only inferred from electronic annotations. In the UniProt database (http://www.uniprot.org) [159] there are currently 148 proteins from *C reinhardtii* with an experimentally validated function, compared to 5,766 for *Arabidopsis thaliana*, and 3,255 for *Escherichia coli*.

The electronic annotation of *C. reinhardtii* is an ongoing process, and so far out of a total of 15,000 proteins there are roughly 7,000 proteins available in the UniProt database that are functionally annotated with at least one GO term. However, when we take the reviewed proteins of *Chlamydomonas* into

account, there are only 299 proteins with a high-quality annotation available.

2.2. The state of microalgal annotations

For microalgae, the inability to obtain a high-quality functional annotation for the majority of the proteins seems to be a returning trend. In Table **1** we bring some recently annotated microalgae and show how deep they are annotated. All of these microalgae were annotated using standard homology-based methods [109, 160-166]. For each of these microalgae roughly half of the proteins lack any form of functional annotation documented in their UniProt database entry [159]. This is likely the direct result of a lack of phylogenetically close well-annotated model species. That does not necessarily mean that the annotations obtained are unspecific or inaccurate, but it does imply that accurate electronic annotations are retrieved only for highly conserved proteins common amongst many microalgae.

2.3. Diversity of Microalgal Annotations

The diversity of microalgae makes them unique biological reservoirs for bioprospecting, and it would be interesting to see how a good quality functional annotation can contribute to this process. By taking the Gene Ontology (GO) annotations from the microalgae species presented in Table **1** into account, and by checking the occurrence of these terms in the nearest well-studied model species *A. thaliana*, we can get hints about the diversity and origin of microalgal proteins annotations (Fig. **1**). The figure shows that 88% of the GO terms assigned to microalgal proteins also occur in *Arabidopsis*. Overall, 85% of the microalgal GO terms are used in the annotation of *C. reinhardtii* protein, but only 7% of the specific *Chlamydomonas* GO terms do not occur in *Arabidopsis*.

For all other species the amount of mapped GO terms is far less than in *C. reinhardtii,* showing an even less diverse annotation. With such little amount of microalgal GO terms, that are not also mapped to *Arabidopsis*, it becomes clear that the current annotation of microalgae largely describes the conserved core-metabolism shared between eukaryotic photosynthetic organisms, and as such will only provide a small contribution to the process of bioprospecting.

To summarize, microalgal experimental protein data is very limited, and due to the large phylogenetic distance to the better-characterized model species large amounts of proteins remain unannotated (Table **1**). To circumvent these bottlenecks, it is necessary to use more advanced annotation methods.

Table 1: Number of computationally annotated and experimentally verified proteins in microalgae and reference species.

Species	Total proteins	Annotated proteins	Experimentally validated proteins
Arabidopsis thaliana	52,178	32,354	5,766
Escherichia coli (strain K12)	6,061	5,399	2,619
Saccharomyces cerevisiae (strain ATCC)	6,729	5,919	4,162
Chlamydomonas reinhardtii	15,152	7,385	148
Auxenochlorella protothecoides	7,193	4,072	0
Bathycoccus sp.	7,889	4,203	0
Chlorella variabilis	9,879	5,409	0
Coccomyxxa subellipsodae	9,802	5,270	0
Micromonas pusilla	10,279	4,782	0
Nannochloropsis gaditana	15,361	7,419	0
Ostreococcus tauri	7,912	4,555	0
Ostreococcus lucimarinus	7,401	4,326	0
Volvox carteri	14,833	6,292	18

*Model species are in bold. Data were taken from UniProt (http://www.uniprot.org) the 1st of August 2015. Total proteins were taken by querying the species taxonomy, total annotated proteins were retrieved by including the query for proteins to have a GO term using any assertion method, and experimentally validated proteins were obtained by only including proteins that have a GO term using any experimental assertion.

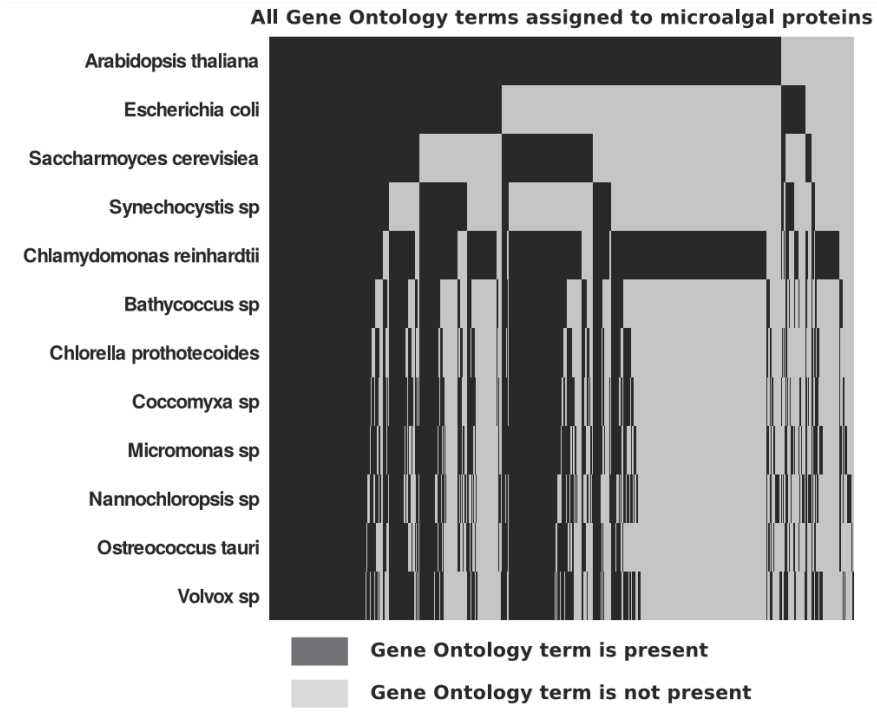


Figure 1: Heatmap showing the presence of microalgal GO term annotations in microalgae and in model species. Y-axis, species presented in Table 1; X-axis, GO terms annotated to microalgae, sorted by their assignment to at least one protein in descending order of the species list.

3. IMPROVED FUNCTIONAL ANNOTATION

3.1. Keeping Up to Date with Annotation Methods - The CAFA Experiment

One way to improve the functional annotation of microalgae, is by applying the latest state-of-the-art tools. The classical sequence similarity-based annotation methods often remain the first method of choice because of their success in the past. However, functional annotation of proteins is a hot topic in the bioinformatics community and new tools are published every year. To keep track of these tools and how well they perform, the Critical Assessment of Protein Function Annotation (CAFA) experiment attempts to rank them according to their performance [56]. The first edition showed that as many as

33 new methods outperformed the standard BLAST-based method. This can be explained by the fact that many of these tools apply sequence similarity-based predictions in different ways, for

example by using one-to-many homology-based annotations, or by using context-aware principles, as described by [157]. An example of a new method that uses a one-to-many approach is Argot2 [2], which combines BLAST results with sequence analysis using profile hidden Markov Models (HMM) and subsequently compares predictions using a semantic similarity approach. An example of a context- aware approach is FFPred2[167], which attempts to transfer functional annotations from known (human) proteins to unknown proteins with similar biophysical attributes.

The CAFA experiment provides a good ranked overview of state-of-the-art annotation tools. However, most of these tools in part still rely on primary sequence similarity, and the structure and context-based alternatives require extensive training sets. Thus, although these tools will most likely produce more reliable results than the classical mainstream functional annotation tools, they will still be unable to annotate many microalgal proteins.

3.2. Consensus-Based Annotation

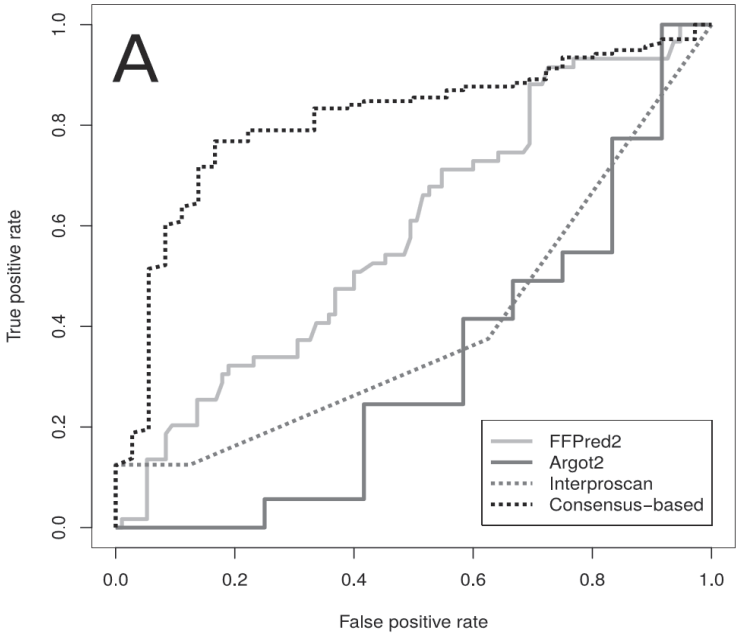
An alternative way to improve the results of protein functional annotation is by using a set of complementary tools and combining individual predictions in a statistical solid manner. For example, by combining FFPred2 with Argot2 we combine a one-to-many homology-based annotation method with a context-aware annotation method. This can be further complemented with a protein domain homology-based transfer of annotation approach using InterProScan [61]. If we would then take the GO term predictions of each of these methods and compare predicted GO terms using a semantic similarity approach as applied in Argot2, we obtain a comparison between predictions of each of these methods, and the specificity for each predicted GO term. By then applying a machine-learning algorithm such as Random Forest we are able to reassess the validity of each of these predictions.

As a test-case we applied this method to a test-set of all new microalgal SwissProt (http://www.uniprot.org) proteins entries between the 1st of July 2014 and 1st of July 2015, using the UniProtKB [159], Uniref90 [168], and Pfam [169] databases from before the 1st of July 2014 as reference. The experiment

was set up with double 10-fold cross validation. Ten data sets were generated with 90% of the predicted GO terms assigned to the training set, and 10% as test-set. The training set was used to train a Random Forest model on the input using 10-fold cross-validation. In this way each final predicted GO term has no influence on the model used to predict these GO term, eliminating overfitting. The accuracy of this method largely improves over that of FFPred2, Argot2, and InterProScan (Fig. **2A**). A test-set of non-algal proteins (Fig. **2B**) was used to compare results with the algal data set. There is a noticeable difference in the performance of FFPred2, Argot2 and InterProScan. For microalgae, the latter two showed lower prediction accuracy.

3.3. Functional Annotation with Hidden Markov Models

Profile Hidden Markov Models (HMM's) provide a statistical description of a sequence family consensus [62]. Effectively a profile HMM turns a multiple sequence alignment of a specific protein family into scoring system that takes into account position-dependent amino acid distributions and position-dependent insertion and deletion gap penalties, which makes this technique suitable for searching remote homologs. To obtain the best model while keeping a high specificity, it is important to build it from experimentally validated proteins only. Because there is no database that contains HMM's built from experimentally validated proteins only, for each specific function a new HMM has to be built, which makes this method not easily applicable for high throughput annotation. Moreover, when selecting only experimentally validated proteins of a specific function, the amount of experimentally verified sequences available can often be too limited.


A
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
FFPred2
Argot2
Interproscan
Consensus−based

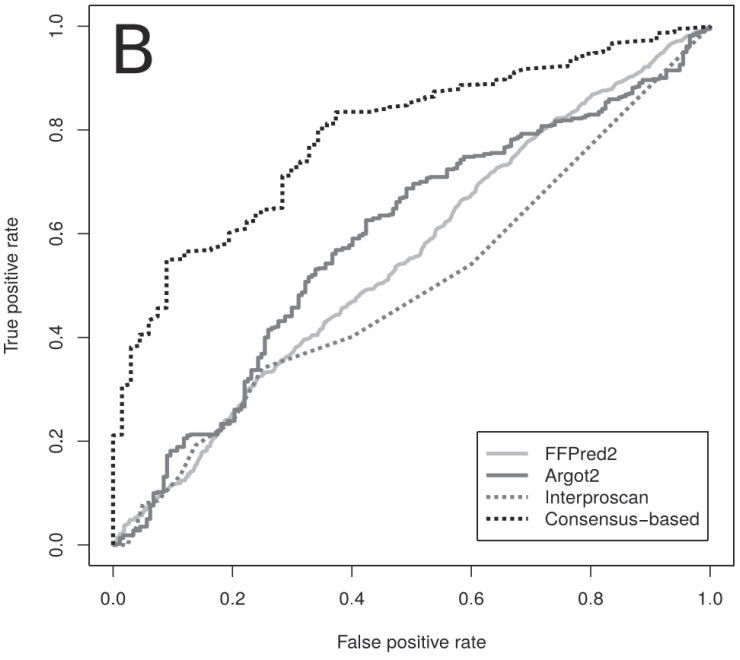

B
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
FFPred2
Argot2
Interproscan
Consensus−based

Figure 2: Receiver Operator Curve of three protein function prediction methods, as well as the consensus-based method that combines these. The test-sets used consisted of protein entries not present in UniProt (http://www.uniprot.org) before the 1st of July 2014. The test-set of (A) 246 reviewed microalgal proteins comprising 540 assigned GO terms, and (B) consisted of 2,429 reviewed proteins comprising 11,701 assigned GO terms. Predictions were done using database versions from before the 1st of July 2014.

3.4. Validation of Matching Proteins *via* 3D Structure

An extension to sequence similarity techniques is 3D structure prediction. Once the best matching proteins are found, they can be used in homology-based modeling approaches. Homology-based modeling uses a reference protein as a base to predict the 3D structure of the protein of interest. The two 3D structures can then be scored for overall quality and be compared. Several tools are capable to estimate the tertiary and quaternary structure of proteins in such a way. One example is SWISS-MODEL (http://www.swissmodel.expasy.org), a web-based tool aimed to provide easy access to predict protein 3D structure from its amino acid sequence, assisted by homology modeling techniques as explained above [170]. Regardless of the automated 3D modeling limitations, scores such as QMEAN, coverage, and identity, can provide an interesting addition to estimate the similarity between the protein of interest and the reference protein.

3.5. Model-assisted annotation

A genome scale metabolic reconstruction aims to integrate biochemical metabolic pathways in a single network and provides a structured platform to correspond metabolic genes with metabolic pathways [171]. As an alternative to laborious manual metabolic model construction, tools such as SEED [172] and Pathway Tools [173] are capable of automatically generating metabolic maps from pathway databases and enzyme annotations. While these tools often provide only a basic overview of an organism's metabolic capabilities, as the topology and breadth of the network is largely dependent on available data, even for microalgae these models can provide valuable insights. For instance, an orphan reaction in metabolic pathways can simply be due to a missed or a too broad annotation. With this information in mind it becomes feasible to use more elaborate, case-by-case, manual annotation methods to close these gaps. A simple first step could be to take the GO term specific for the particular protein and link this to similar but more generic parental GO term(s). Proteins annotated with these less specific GO term(s) are then considered to be promising candidates for the missing GO term and should be re-evaluated. One way to do this is by building an HMM based on UniProt proteins that are experimentally validated to have the specific GO term. This statistical model can then be applied to the candidate protein selection, and in this way, we might be able to identify the correct protein.

3.6. Subcellular Localization

Protein localization prediction is an important part of a protein's functional annotation. If two proteins involved in the same reaction are functionally assigned to a different subcellular compartment, the reaction cannot occur. On the other hand, microalgae are known to possess multiple iso- functional proteins that essentially perform the same reaction but in different subcellular compartments. This information is crucial for the more elaborate compartmentalized genome scale models of microalgae. One robust way of figuring out the subcellular location of (isofunctional) proteins is by performing subcellular proteomics [174], but this is often technically difficult, expensive, and time consuming. The UniProt database currently contains 529 reviewed microalgal proteins with a subcellular location annotated, of which 54 are experimentally validated. Therefore, it is necessary to computationally predict the subcellular location of proteins. For this purpose, several tools are available, such as: Argot2 [2], TargetP [95], SignalP[96], PSORTb [97], and PredAlgo [98]. However, with the exception of PredAlgo, most of these tools are trained with different types of species in mind, resulting in predictions that do not take into account the specific cellular arrangements and compartments in microalgal species.

PredAlgo is a predictor specifically trained for microalgae, using a *C. reinhardtii* based training set of 79 chloroplast, 39 mitochondrial, 39 secretory pathway, and 89 cytosol proteins. It shows good prediction results for *Chlamydomonas* proteins and closely related microalgal proteins. However, for other more distantly related microalgal species predicting subcellular localization is difficult due to their polyphyletic nature. It is believed that different endosymbiotic events happened in parallel, forming the first microalgae [175]. This caused a difference in the arrangements of cellular compartments, or even different types of cellular compartments. Therefore, PredAlgo may not be accurate in predicting protein locations for microalgae not related to *C. reinhardtii*. To circumvent this, the PredAlgo algorithm will have to be trained with proteins from additional microalgal clades. Alternatively, results from multiple predictors possibly can be combined as described above. Finally, it should be noted that PredAlgo only predicts to which compartment a protein is targeted. If the compartment where translation of the protein takes place is unknown it is still not possible to know the final location, or to which membrane it is targeted. Therefore, compart- mentalized omics data is needed to accurately predict the final subcellular location of a protein. The UniProt

database contains 8 chlorophyta proteome sets based on genome sequencing data that are fully annotated. Additionally, GenBank (http://www.ncbi.nlm.nih.gov/genbank/) harbours many chloroplast and mitochondrial datasets that can be used for this purpose [97]. These datasets could be compared to assess the potential impact of the parallel endosymbiotic events, which in turn can be taken into account in cellular location predictions.

In conclusion by using consensus-based protein function prediction, and/or model-assisted annotations, many improvements can be made in functional annotation of microalgal proteins. Nevertheless, there will always be a set of species-specific proteins that will remain unannotated using computational methods.

4. REMAINING BOTTLENECKS

4.1. Unsupervised Computational Annotations Can Lead to Error Propagations

The GO project has become the standard way of annotating proteins [65]. All major databases use these terms for the documentation of protein functions and cellular locations. GO terms are accompanied with an evidence code, stating how a gene was assigned to a GO term [176]. In most cases the evidence code is "Inferred from Electronic Annotation", meaning that an unsupervised computational method was used. Such annotations are error-prone. Furthermore, with an exponential-increasing amount of sequencing data being generated every day, the amount of unsupervised electronically assigned GO terms also increases exponentially. To illustrate the problem, *Schaid et al.* showed that already in 2010, 50% of the 200,000 human protein GO term assignments were done electronically. Consequently, these gene annotations were likely to contain a number of errors [177]. If such an electronic annotation is done using, for example, a standard BLAST based transfer of annotation method and proteins that also have their GO terms electronically assigned were used as a reference, this can easily lead to error propagation of GO term assignments. Recording the provenance of unsupervised annotations to GO terms is therefore essential. Several annotation tools are scoring GO term annotations based on the GO structure and evidence codes [178]. The evidence ontology (ECO) [176] provides more descriptive evidence-based annotation to proteins in UniProt database by describing, for example, evidence types, methods and data curation. A detailed provenance can help to obtain more precise evidence scores than is possible with the standard evidence codes.

4.2. Orphan Reactions

One of the direct results of a small amount of experimentally validated proteins is a large amount of orphan reactions. These are reactions catalysed by enzymes of which it is assumed that they must occur, for instance through phenotypic analysis or because they bridge a gap in a metabolic pathway, but which do not have an encoding gene assigned. The Orphan Enzymes Project (http://www.or phanenzymes.org) is attempting to link sequences to such Orphan Enzymes, and a similar effort should be made regarding microalgal enzymes.

4.3. The Lack of Identified Microalgal Specific Protein Domains

An effective way of assigning GO terms to proteins is by linking GO terms to protein domains and searching proteins for the presence of these domains. However, as can be seen in Fig. (**2B**) (InterProScan results) microalgae proteins show a low level of sequence similarity to domains available in the PFAM database (http://www.pfam.xfam.org), which suggests that microalgae have accumulated many novel domains that are not yet identified by the scientific community. To start to identify these novel domains it might be useful to develop an algae specific domainome by routinely performing large-scale comparative genomics between all available microalgal genomics data, as was done with bacteria [179]. Recurring patterns can then be assigned to specific domains with presently unknown function. If specific domains keep recurring in proteins associated with specific traits, these domains can be linked to a function.

CONCLUDING REMARKS

Systems biology approaches to unlock the potential power of microalgae are seriously hampered by lack of genomic knowledge. Genome annotations of recently sequenced species still heavily depend on sequence- similarity based functional annotation methods, which are less suitable for species that have no close by well-studied and annotated homologous species. As a result, almost 90% of microalgal functional protein annotations is still for the most part describing the metabolic core shared between algae and plant species. The application of novel state-of- the-art annotation methods, as well as approaches that combine multiple methods, may result in a more accurate and more diverse functional annotation. Genome scale modeling approaches could additionally help in identifying metabolic gaps, which can then be looked at more thoroughly. However, for microalgae to fulfil their promise as a

biosynthetic host it is important to overcome at least some of the annotation bottlenecks that are not solvable by computational methods. We therefore suggest that a large- scale wet lab effort focused on a number of selected microalgal reference species is essential. This would provide the computational methods with larger, more diverse set of reference genes, and would allow computational annotations methods to quickly tap into the promising biological reservoirs of industrially interesting algal species [180].

Chapter 4

CrowdGO: a wisdom of the crowd-based Gene Ontology annotation tool

Maarten J.M.F. Reijnders, Vitor Martins dos Santos and Peter J. Schaap

Laboratory of Systems and Synthetic Biology, Wageningen University, Wageningen, The Netherlands

Manuscript in preparation

Abstract

Motivation: De novo protein function prediction has been a hot topic in bioinformatics since the early days, and even more so since the start of the omics era. However, predicting protein functions in high-throughput is notoriously challenging. Most prediction methods are based on sequence similarity or on machine learning. Sequence similarity-based annotation methods often have a high specificity and sensitivity in case of well-characterized orthologs but are unable to predict any functionality when these are absent. Homology-independent machine learning based methods do exist but usually have a lower specificity, as other protein features are less informative than sequence conservation. In an ideal case scenario specificity and sensitivity should be combined.

Results: To achieve a higher sensitivity and specificity in de novo protein function prediction, CrowdGO combines multiple homology-dependent and independent protein function prediction methods. It uses Gene Ontology semantic similarity to correlate and compare the various functional predictions and reassesses the predicted terms using a random forest algorithm. Based on a test set, CrowdGO shows a significant area under the curve increase when assessing sensitivity and specificity. This is also showcased by a net-gain in true positive and true negative predictions.

Conclusion: Given the significant increase in both sensitivity and specificity, CrowdGO would be a good addition to any omics study in need of high-throughput prediction of the encoded functionome.

Availability: CrowdGO can be found at https://gitlab.com/mreijnders/CrowdGO

1. Introduction

Non-model species are often interesting targets for biotechnological research, but our biochemical understanding of their protein functions is limited. For example, microalgae are almost exclusively annotated based on computational predictions [157]. Standardly used function prediction tools use sequence similarity, where in the case of sufficient sequence homology with a protein its function is transferred over [2, 63]. These methods work well for proteins that show a high level of sequence conservation, by transferring the annotation information from closely related well-studied model organisms. However, it works less well for non-model species, especially when there is no closely related well-studied model organism available. Alternatively, there are homology independent machine learning-based methods, which attempt to correlate protein features such as hydrophobicity, protein domains, and presence of signal peptides, with a protein function [167, 181]. These methods have a high recall but lack precision, as correlating generic protein features to a protein function leads to more ambiguity than transferring function between proteins with a high sequence similarity. Most methods fall under these two categories as can be seen in the CAFA challenges for protein function prediction [56, 182], a competition between scientists where they attempt to predict the function for novel protein sequences as accurate as possible. However, most methods suffer from drawbacks, and ideally, we may want to combine the advantages of multiple methods while negating their drawbacks.

Combining different methodologies to achieve better prediction results is not a new idea. In 2009, Rentzsch and Orengo discussed the new era of genomics, and the advantages that come with it [92]. In their review, they provide a comprehensive overview of function prediction methods, and which biological aspects they use to predict protein functions based on Gene Ontology (GO) terms, a framework of notations used to describe a proteins function. They argued that in the 'age of multiplicity', only the use of multiple tools, multiple evidence, and the multiple aspects of function, can give us a good insight into protein functions. Since then, many tools have been designed that combine multiple biological features to make a function prediction. However, while they argued that multiple tools need to be used to provide a comprehensive insight into protein function, there is currently no standard method to merge predictions of orthogonal different tools.

For this purpose we have designed CrowdGO: a wisdom of crowd based annotation tool that uses GO term semantic similarity and a random forest

algorithm [183] to combine the predictions of multiple methods. CrowdGO is able to use the results of any GO term prediction method that provides confidence intervals. In this paper we use an example based on FFPred2 [167] and Argot2 [2], which were the top performers in the CAFA2 challenge [56], and InterProScan [63]. These three methods were chosen because of their prediction performance, and because they use complementary approaches. FFPred2 is a machine learning based prediction method that uses support vector machines [184] on a set of 13 protein features such as amino acid composition, low complexity regions, and secondary structures. Argot2 is based on both sequence similarity and machine learning, using BLAST [154] and HMMER [185] to calculate similarity to existing protein sequences, and a similarity calculation between the BLAST and HMMER retrieved GO terms. Finally, InterProScan uses machine learning by training profile Hidden Markov Models [62] to predict protein domains, and transfers any GO terms associated to these domains to the protein.

2. Methods

The CrowdGO annotation pipeline consists of two parts: correlation of gene ontology (GO) terms [64, 186] generated by the different prediction tools, followed by a random forest algorithm [183] to distinguish between true positive and false positive predictions.

2.1 CrowdGO workflow

Figure 1 visualizes a simplified workflow of the CrowdGO annotation process:

1. The user selects their proteins for analysis, and a training set consisting of proteins with a known function.
2. All proteins are initially assigned GO terms by two or more prediction methods. In our test case we used FFPred2 [167], InterProScan [63], and a reversed engineered local implementation of Argot2 [2].
3. GO-term assignments from each method are compared. Protein-GO term predictions that are in the same GO hierarchy get clustered with given similarity scores. The GO term with the highest IC score (Equation 1) is chosen as representative term.
4. All scores from step two and three are entered in a random forest model [183] (Table 1). The random forest is trained on the training set using 10-fold cross-validation and is used to predict whether a previously predicted GO term is a true or false positive.
5. The output of the pipeline is a list of protein-GO term pairs with a confidence interval between zero and one.

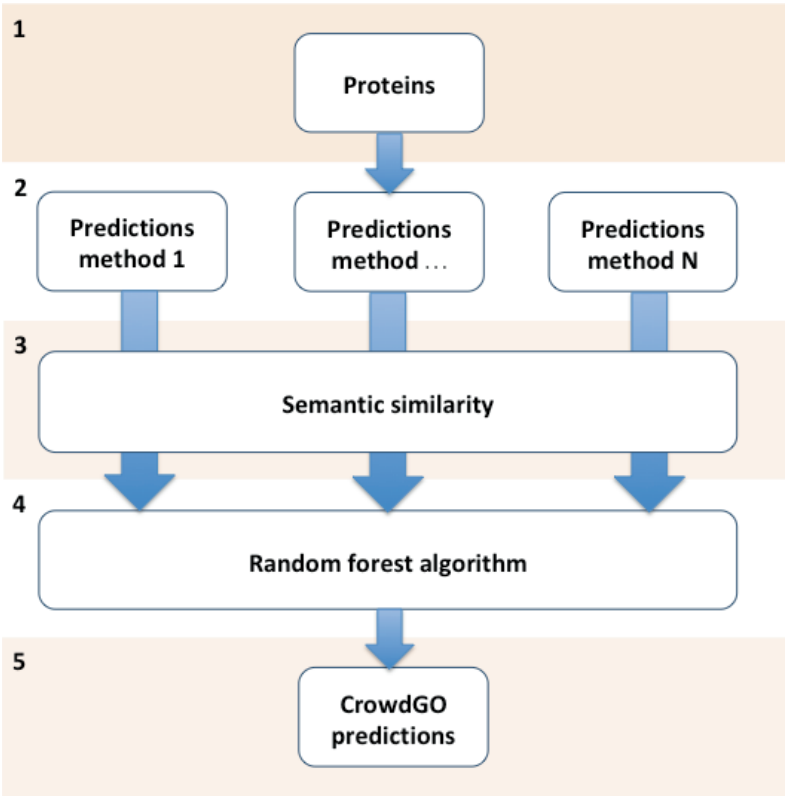


Figure 1: Simplified workflow of the CrowdGO annotation process. 1) A protein with an unknown function is 2) annotated by two or more existing methods. 3) The similarities for the predicted GO terms between each method are calculated (Equation 1,2). 4) All the scores produced by step 2 and 3 are used as an input for a random forest algorithm. 5) The outcome of the random forest algorithm gives confidence intervals between zero and one for each previously predicted GO term.

2.2 Methods used for GO term prediction in step two

For our testing of CrowdGO we used three different predictors for step two: a reversed engineered local version of Argot2 based on [2], FFPred2 [167], and InterProScan [63]. For training purposes of these tools, we only used databases and other prior information from before 01-07-2015, to avoid biased predictions and over fitting. The versions can be found in table 1.

2.3 Usage of Argot2, FFPred2, and InterProScan

Our implementation of Argot2 (ArgotToo) is made using the published methods [2], as there was no local implementation publicly available. ArgotToo uses BLAST [60] and HMMScan [185] results as an input. BLAST input was generated by a BLASTP on the UniProt Knowledgebase database from 24-06-2015 (Table 1) [55]. The HMMScan input was generated on PFAM version 27 (Table 1) [169].

For FFPred2 we downloaded the available local version (http://bioinfadmin.cs.ucl.ac.uk/downloads/ffpred/) and used its default settings. After calculating the optimal cut-off for precision and recall using pROC [187], the true positive threshold was adjusted to 0.7 instead of the default 0.5. This threshold was used in evaluating FFPred2 in Table 2.

We used InterProScan version 5.13-52 and disabled its pre-calculated lookup service. Since InterProScan provides no unified scoring system for each of its sub-programs, we used the amount of times a GO term was predicted to a protein independently. This score ranges from 1 to 26, which was the maximum amount of times one GO term got predicted to a protein. However, we chose to only use this number as an input for the random forest model and used a prediction threshold of 1 for evaluating InterProScan in table 3.

Table 1: Database versions used in the prediction of GO terms

Database	Version	Date
UniProt SwissProt	2015_06	24-06-2015
UniProt TrEMBL	2015_06	24-06-2015
UniProt Uniref90	2015_06	24-06-2015
Pfam A	27	22-05-2013
InterPro	52	22-05-2015
Gene Ontology Annotation	144	23-06-2015

2.4 Calculating the semantic similarity between GO terms

The gene ontology comparison algorithm is an implementation of Lin's algorithm [188], as similarly used by Argot2 [2]. For each GO term pair between two prediction methods we compute the similarity score and their information content (IC) score. The IC score is used to assess how common the GO term is in the UniProt Knowledgebase. A GO term its IC score is calculated as follows:

(1)

Where GO:GO_i is the total number of times GO term $_i$ is represented in the UniProt database, and GO:GO is the total number of GO terms assigned to all proteins in the UniProt database.

The similarity score is then calculated as follows:

(2)

Where icParent(GO_i,GO_j) is the parent term shared between GO_i and GO_j with the highest IC score, or the highest IC score between GO_i and GO_j if one is a parent term of the other.

2.5 Labelling the predictions for training the model in step four

In step 4 of Figure 1, CrowdGO trains a random forest model that requires a set of proteins with known GO terms. Predictions for these proteins are compared to these known GO terms, and subsequently labelled true or false positive based on the Gene Ontology (GOA) hierarchy [65]. If the predicted GO term is the same as the real term, or in the same GOA hierarchical structure

excluding its root term, it is labelled as a true positive. Otherwise it is labelled as a false positive. The same labelling is used to evaluate the test set predictions, in addition to true negatives and false negatives. A true negative is a GO term that is correctly not annotated to a protein by being below the methods confidence threshold, and a false negative is a GO term that is incorrectly not annotated by being below the methods confidence threshold.

2.6 ROC plot calculations.

The calculations and drawing of figure 2 were done using the R package pROC [187], and the calculations and drawing of figure 3 were produced using the R package PRROC [189].

Table 2: All the data we have of the predicted GO terms, using all input prediction tools. The data used for the Random Forest algorithm is indicated in the 'RF' input column. * The classifier is only given in case of training the model

Data	RF Input
Protein identifier	No
GO term	No
GO term IC score	Yes
Number of predictions in cluster	Yes
Argot2 GO term	No
Argot2 score	Yes
Argot2 GO IC score	Yes
FFPred2 GO term	No
FFPred2 score	Yes
FFPred2 GO IC score	Yes
IPRScan GO term	No
IPRScan GO term score	Yes
IPRScan GO term IC score	Yes
Argot2 - FFPred2 GO similarity	Yes
Agrot2 - IPRScan GO similarity	Yes
FFPred2 - IPRScan GO similarity	Yes
GO root term (BP/MF/CC)	Yes
Classifier (True/False)	Yes*

3. Results

3.1 Local implementation of Argot2

We wanted to use Argot2 as one of the three methods in this paper to evaluate CrowdGO. However, there is no local version available for Argot2, which we need for the blind predictions for the training and evaluation of our proteins. Therefore, we recreated a local version of Argot2 based on the paper, called ArgotToo. We took 250 random SwissProt entries created in 2017 to evaluate their performance. For this evaluation we used the same databases for ArgotToo as for Argot2, which are all from before 2017. The result is shown in Figure 2. Small deviations are likely caused due to differences in database handling, but in general the predictions are the same.

ArgotToo can be found at: https://gitlab.com/mreijnders/ArgotToo.

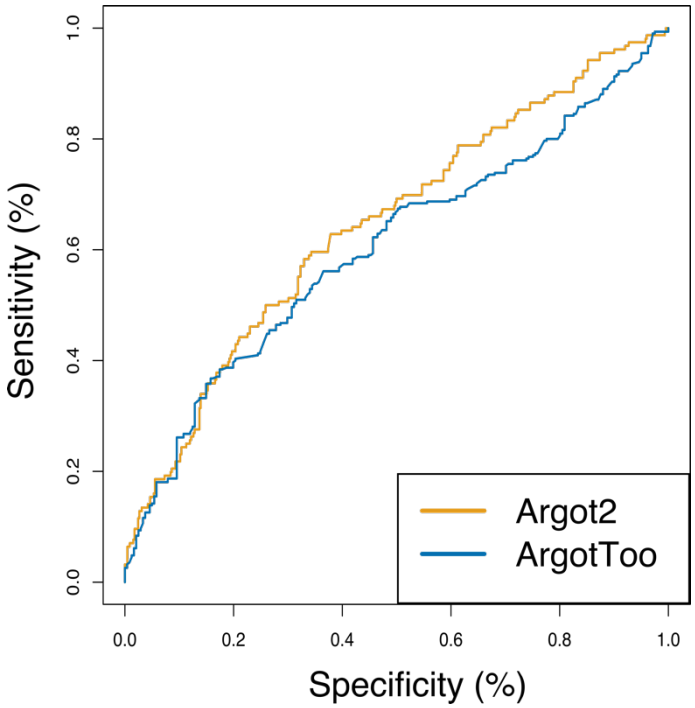


Figure 2: Area under the roc curves for Argot2 and ArgotToo. On the x-axis the false positive rate, and on the y-axis the true positive rate.

3.2 Training and evaluating the random forest model

The random forest model is trained on GO term predictions of 3398 proteins. Here we selected all new SwissProt [55] proteins between 01-07-2015 and 01-07-2017 for which there is one or more experimentally validated GO term available. A thousand proteins were randomly selected for a test set, and the remaining 2398 proteins were selected for the training set. Before training the random forest, the input GO term predictions were reduced and equalized between true and false positives to remove any potential bias. The exact input used for our test case can be found in Table 2.

3.3 Assessing sensitivity and specificity

In Figure 3 the sensitivity and specificity of each method is compared. The ROC curve shows predictions made by CrowdGO, ArgotToo [2], FFPred2 [167], and InterProScan [63]. The CrowdGO cut-off score to achieve the maximum combined precision and recall is marked at 0.6. Further, we compare methods for only biological processes, molecular functions, and cellular components type of functions (Figure 3 B, C, D). Notable is the lack of all methods to predict cellular components (Figure 3 D), and a slight increase in prediction power for ArgotToo for molecular functions (Figure 3 C). Figure 4 shows the comparison for each method its precision-recall curve.

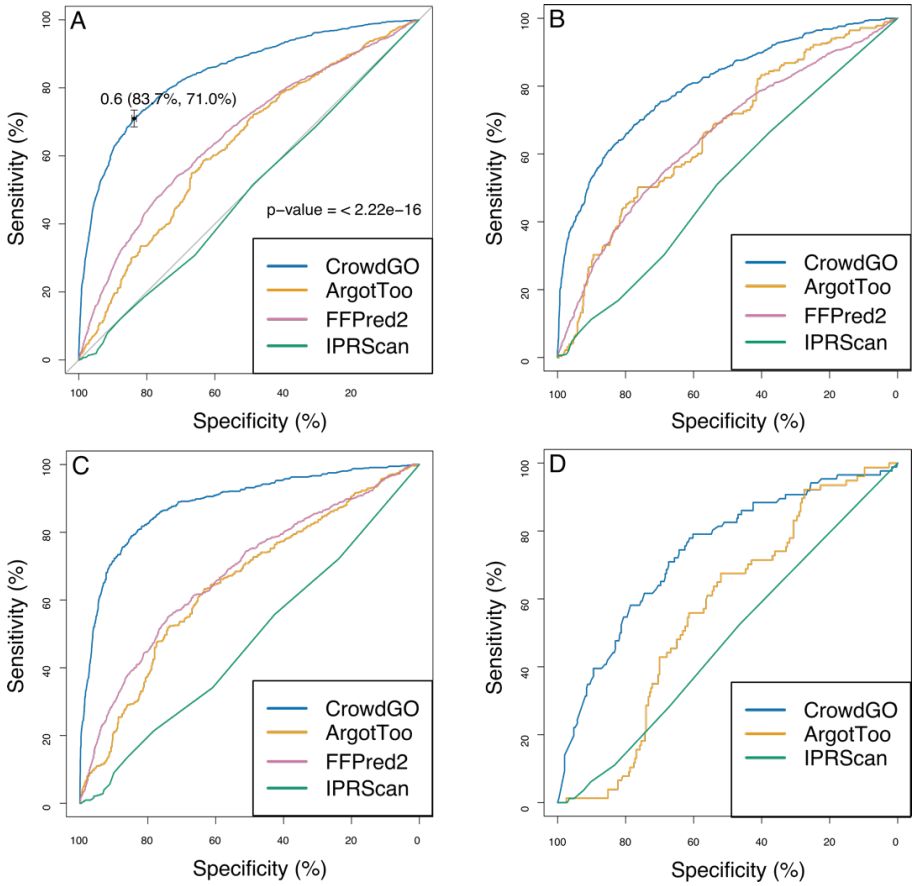


Figure 3: A) Receiver Operator Curve for the various methods, with specificity on the X-axis and sensitivity on the Y-axis. The most optimal sensitivity and specificity combination is calculated to be at the 0.6 threshold, with 84% specificity and 71% sensitivity. All curves compared to CrowdGO have a p-value of < 2.22e-16. B) The same ROC curve for only biological processes, C) for only molecular functions, and D) for only cellular components. Note that FFPred2 is omitted from the cellular components because it does not predict these.

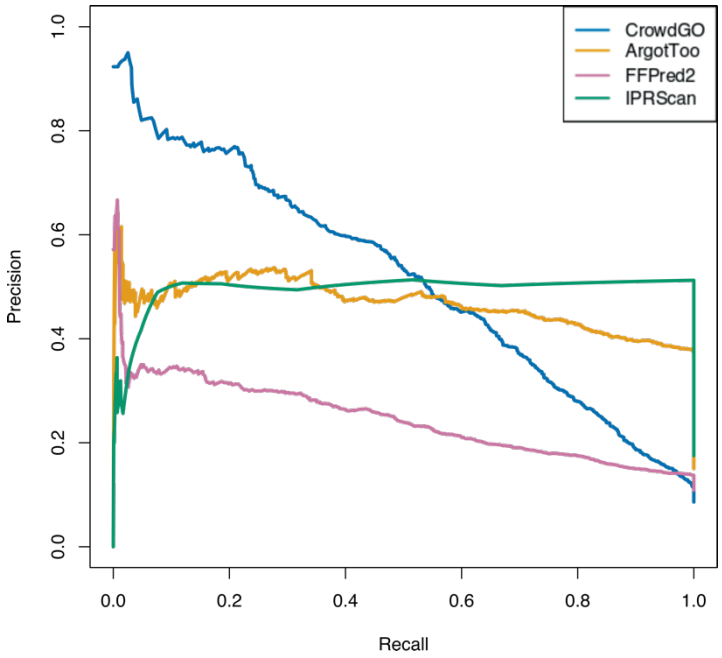


Figure 4: Precision - recall ROC curve comparing CrowdGO to Argot2, FFPred2, and InterProScan.

3.4 Observing how CrowdGO re-classifies existing input predictions

CrowdGO is specifically designed to reclassify the predictions based on integration between methods. With that in mind we want to observe the effect CrowdGO has on the original GO term annotations. In table 3 we set out the predictions of CrowdGO compared to its original GO term input. E.g., we observe how many true positive predictions became a false negative, and how many false negatives became a true positive.

Table 3: Amount of predictions that changed from true positive (TP), false positive (FP), true negative (TN), false negative (FN), or stayed the same after being assessed by CrowdGO. Before the arrow: the GO term as assessed by each original method. After the arrow: the GO terms as assessed by CrowdGO.

Method	Tp->Tp	Tp->Fn	Fp->Fp	Fp->Tn	Tn->Tn	Tn->Fp	Fn->Fn	Fn->Tp
Argot2	556	46	447	365	817	323	87	191
FFPred2	222	77	349	1.427	7.961	841	298	385
IPrScan	305	35	232	328	0	0	0	0
All	1.083	158	1.028	2.120	8.778	1.164	385	577

4. Discussion

In the paper we introduced CrowdGO, a protein function prediction tool which utilizes existing predictions and reassesses them using GO term semantic similarity and machine learning. One major selling point for CrowdGO is its ability to handle all types of input data. In the case of FFPred2, for which the raw results seem to include a lot of false positives, CrowdGO is able to reclassify a lot of these false positives to true negatives while only a few true positives are reclassified to false negatives.

InterProScan does not give us any meaningful confidence scores for its predictions, which means their sensitivity-specificity curve (Figure 3) and precision-recall curve (Figure 4) are meaningless. Also, the lack of meaningful confidence intervals means we were unable to classify predictions as negatives (Table 3). With CrowdGO we are able to combine the InterProScan predictions with other predictions to not only improve the results (Table 3), but also attach meaningful confidence intervals to the predicted GO terms.

When evaluating the isolated ArgotToo results in table 3, we notice its results are the least affected by CrowdGO. This is likely due to Argot2 having relatively reliable predictions. However, Argot2 provides a lower amount of total predictions compared to FFPred2 and InterProScan (Table 3), with the exception of the GO category Cellular Component. In combination with tools that predict more GO terms but are arguably more prone to false positive predictions, CrowdGO is able to reliably predict more GO terms than ArgotToo could do by itself (Table 3).

Both the biggest upside and downside of CrowdGO is its heavy reliance on the input predictions. If these are of low quality, the CrowdGO results will not be much better; if these are of high quality, the CrowdGO results will be of even higher quality. Additionally, combining input predictions from complementary techniques such as sequence similarity and machine learning would potentially enhance the performance of CrowdGO. Certain proteins might be hard to annotate by one technique, but easier to annotate by a different technique. Given a proper training set for CrowdGO, the random forest model would be able to recognize patterns where the predictions of one technique would be of more value than that of another technique. All of this requires a good understanding of protein function prediction techniques, resulting in a good input prediction set. Therefore, while CrowdGO shows a significant improvement in predicting protein functions, it takes a basic

knowledge of protein function prediction to be used effectively. One way of making CrowdGO widely usable by the community, is by incorporating it in existing protein function annotation pipelines. It would be wise to test multiple existing prediction methods combinations together with CrowdGO, to see which tools and combinations deliver the best results. This way CrowdGO can be distributed as part of a protein function prediction suite, with easy-to-use instructions usable by everyone.

5. Conclusion

We have successfully improved GO term predictions of existing methods by combining their results using CrowdGO. In particular, CrowdGO shows significant improvement in sorting high-confidence predictions from low-confidence predictions, resulting significantly higher area under the curves (Figure 3), and a better precision-recall curve (Figure 4). Furthermore, CrowdGO is able to reassess false positives and false negatives to true negatives and true positives respectively (Table 3).

Chapter 5

Genome-scale annotation of the oleaginous yeast *Cutaneotrichosporon curvatus* ATCC 20509 using CrowdGO

Maarten J.M.F. Reijnders[1], Nhung Pham[1], Stefan Aanstoot[2], Gerrit Eggink[2], Jan Springer[2], and Peter J. Schaap[1]

[1]*Laboratory of Systems and Synthetic Biology, Wageningen University, Wageningen, The Netherlands*

[2] *Bioprocess engineering, Wageningen University, Wageningen, The Netherlands*

Manuscript in preparation

Abstract

Motivation: Yeasts are frequently used for biotechnological applications, however predominantly from the ascomycota subdivision. Yeasts from the basiomyceta subdivision are underrepresented despite their interesting characteristics such as producing high amounts of triterpenoids, carotenoids, and complex carbohydrates. Because these species are underrepresented, we know relatively little about their genome organization. In this study we aim to functionally annotate the genome of the basiomyceta *Cutaneotrichosporon curvatus* using CrowdGO from **chapter 4** and compare the results with manual functional annotation of metabolic proteins to assess CrowdGO in a real-case scenario.

Results: We compared the CrowdGO annotations of *Cutaneotrichosporon curvatus* against the existing annotations of the related *Cutaneotrichosporon oleaginosum* and show a vast increase in amount of proteins retrieving GO term annotations. GO enrichment analysis using differential expression analysis of nitrogen and non-nitrogen growth conditions showed an enrichment of GO terms that would be expected from an oleaginous species. To increase the resolution of functional descriptions CrowdGO was implemented in a manual annotation pipeline to manually curate 700 metabolic proteins. Together with a differential expression analysis, these functional annotations were used to characterize triacylglycerol synthesis of *Cutaneotrichosporon curvatus*. Both the CrowdGO high-throughput annotations, and its utilization in a manual annotation pipeline, show promise towards the improvement of annotations for oleaginous yeasts.

1. Introduction

1.1 Basidiomycota yeasts and biotechnology

Yeasts are the species with the most biotechnological applications, in particular the model species *Saccharomyces cerevisiae* of the ascomycota subdivision [190]. The ascomycota yeasts are known for applications in fermenting food and drinks, heterologous protein production, probiotics, and many more. These yeasts are well studied because of their applicability, and because of their potential in biotechnology, they are studied in even more detail. The subdivision basidiomycota, however, is not widely used in biotechnological applications, despite some interesting characteristics amongst which the most promising is their potential to produce large amounts of secondary metabolites such as terpenoids, carotenoids, and complex carbohydrates. To try and break this circle, it is interesting to study functional genomics of basidiomycota species.

1.2 The oleaginous basidiomycota yeast *Cutaneotrichosporon curvatus*

Cutaneotrichosporon curvatus is a non-model oleaginous yeast of the subdivision basidiomycota, able to produce up to 60% triacylglycerol (TAG) of its dry weight [191]. It is able to grow on low-cost carbon sources such as whey permeate, molasses, and other sugar wastes [192, 193]. These characteristics made *C. curvatus* an interesting organism to study since the 80's [192]. Studying its TAG production is mostly focused on the dynamics of nitrogen starvation: an insufficient amount of nitrogen leads to an increase in TAG synthesis, and lowers the cells growth rate [191]. While the general TAG synthesis pathway in yeast is well understood, predicting which proteins are involved in each step is challenging [56]. Furthermore, given the challenges in protein function prediction, automated annotations are often incomplete (**chapter 3**)[194], potentially leaving out key enzymes and transporters. Accurately annotating the metabolism of *Cutaneotrichosporon curvatus*, and in particular its TAG production in relation to nitrogen levels, would give valuable insights for any potential biotechnological applications for this yeast and can act as a reference for other basidiomycetes.

1.3 Functional genomics of Cutaneotrichosporon curvatus

While model species generally have a large amount of their protein functions assessed in wet-lab experiments, non-model species such as *C. curvatus* do not. As a result, large-scale computational predictions need to be done to retrieve their protein functions. *Cutaneotrichosporon oleaginosum* is a species closely related to *Cutaneotrichosporon curvatus*, with over 8,500 proteins in the TrEMBL UniProt database, and for which a bit more than half have one or more GO terms assigned to it using UniProt's UniRule and SAAS annotation pipelines [159]. These pipelines attempt to find protein domains and other functional regions and annotate Gene Ontology (GO) terms to proteins using a manually (UniRule) or computationally (SAAS) generated set of rules, based on for example domain presence - absence, or taxonomic evidence.

However, as discussed in **chapter 3 and 4**, annotating non-model species in high-throughput can be particularly challenging. The estimate for the age of the Ascomycota and basidiomycota split was between 1 and 2 GA ago [195]. Annotating GO terms of non-model species using only sequence similarity to reference proteins or domains will likely be incomplete at such a large phylogenetic distance.

Chapter 4 addresses this issue with the introduction of CrowdGO, a protein function prediction tool that merges and improves Gene Ontology (GO) term annotations from other high-throughput prediction tools. In this study, we use CrowdGO GO term annotations to study the metabolism of *C. curvatus*. However, CrowdGO is tested on an artificial data set that does not represent a real functional genomics study. Therefore, the functional genomics study of *C. curvatus* can act as a hands-on scenario to assess the performance of CrowdGO.

1.4 Manual curation of *Cutaneotrichosporon curvatus* metabolism

For the further assessment of CrowdGO, and to gain high-resolution GO terms in *C. curvatus*, we use CrowdGO as part of a biocuration pipeline of the *C. curvatus* metabolic proteins. This was done using CrowdGO annotations as a starting point, and comparative genomics with the manually curated metabolic proteins of the ascomycota yeast *Yarrowia lipolytica*. By comparing the electronically inferred CrowdGO predictions to those of the biocurated proteins we can assess the performance of CrowdGO.

1.5 Differential expression of *Cutaneotrichosporon curvatus* to study triacylglycerol synthesis

Finally, we performed differential expression analysis using our transcriptomes of nitrogen replete and nitrogen deplete growth conditions and used this to further characterize the TAG metabolism of *C. curvatus*. This detailed metabolic map allows us to generate hypotheses regarding the genetics of TAG synthesis during nitrogen starvation, and to assess the accuracy of the manual annotation.

1.6 Aim

The main aim of this chapter is to assess the performance of CrowdGO in a real-case scenario, while a sub-aim is to perform a functional genomics study on *Cutaneotrichosporon curvatus* to improve our knowledge of basidiomycetes metabolism, with a special interest for its triacylglycerol synthesis. CrowdGO is assessed throughout each step of the functional genomics study, either by comparing it to the annotations of an existing basidiomycota, assessing GO enrichment analysis, or by comparing it to the biocurated proteins.

2. Methods

2.1 Culturing of Cutaneotrichosporon curvatus

Cutaneotrichosporon curvatus ATCC 20509 was selected for culturing. Two growth media were used based on Meesters et al 1996 [191]. The glycerol and NH_4Cl concentrations were adapted to generate our desired carbon and nitrogen ratio's (table 1). The carbon nitrogen ratios are taken from Ykema *et al* [196], which shows a *C. curvatus* growth for a ratio of less than 5, and lipid production for a ratio between 20 and 40 carbon / nitrogen.

Table 1: Glycerol and NH_4Cl levels for the different media, as well as the carbon and nitrogen ratios.

	Glycerol	NH_4Cl	Carbon / nitrogen ratio (mol)
Medium A	16	1	28
Medium B	8	5	2.8

C. curvatus was inoculated from a freshly prepared YPD-agar plate in 50 ml of YPD medium and grown O/N in a 100 ml Erlemeyer flask at 30 °C and 225 rpm.

The culture was divided in two 25 ml portions and centrifuged (10 min. 300 rpm) to collect the cells. The cell pellets were resuspended in 30 ml medium A or medium B. 4 ml of the resuspended cells was used to start duplicate cultures in medium A and B which were incubated for 18 hours at 30 °C and 225 rpm. Each culture was divided in two equal portions and the cells were harvested by centrifugation and the wet pellet frozen in liquid and used for RNA extraction, fatty acid analysis and dry weight determination. Medium samples were taken for glycerol and nitrogen analysis.

2.2 RNA extraction procedure

RNA was extracted using an acidic hot phenol extraction procedure. Briefly, the cell pellet was ground in liquid nitrogen and mixed with 4 volumes of pre-warmed (60°C) phenol + extraction buffer (1% SDS, 10 mM EDTA, 0,2 M NaAC (pH 5) after these 2 volumes of chloroform were added and mixed thoroughly. After centrifugation the buffer layer was washed once with chloroform. RNA was precipitated from the buffer layer by adding 8 M LiCL to and end

concentration of 2M. After centrifugation the pellet was washed once with 2M LiCl and twice with 70% ethanol. The remaining pellet was resuspended in RNase free water.

Total RNA extract, RNA sequencing, and RNAseq data processing were performed as described in [197]. Samples were sequenced by NovoGene using Total RNA.

2.3 Proteome comparison to Cutaneotrichosporon oleaginous

We compared our predicted proteins to the existing *C. oleaginosum* proteins in UniProt. All *C. oleaginous* proteins were extracted from SwissProt and Trembl version 2017_12 [55], and used as a BLAST [60] database for our *C. curvatus* proteins. The BLAST hits for each *C. curvatus* protein against a *C. oleaginous* protein were concatenated. If the concatenated BLAST hit length was 99% or more than that of the *C. oleaginosum* protein its length and vice versa, and the concatenated BLAST hit shared 99% or more amino acid identity, we considered it the same protein. If those numbers were 30% or higher but lower than 99%, we considered it an incomplete protein match. In any other case, we considered it a dissimilar protein. All matches were categorized in the following groups: proteins that have a match in both species, proteins that have an incomplete match in one of the species, and proteins that are unique to either one of the species.

2.4 Protein function prediction

Protein function prediction was done using CrowdGO described in Chapter 4. Instead of the training set used in Chapter 4, we created a training set consisting of only fungal proteins created between 01-01-2015 and 01-01-2017. All predictions for the training set were done on database and program versions before 01-01-2015. In the final annotation, all GO terms that were not annotated to any existing fungal protein in UniProt version 2017_01 were removed. We used a cut-off score of 0.6 to differentiate between true and false predictions.

2.5 Biocuration of *C. curvatus* protein functions

We used CrowdGO annotations in conjunction with other methods and visual inspection to manually annotate the metabolic proteins of *Cutaneotrichosporon curvatus*. A simplified overview of the biocuration approach is given in Figure 1.

1) We extracted single proteins from the manually curated *Yarrowia lipolytica* genome-scale model iNL850 [198]. The protein sequence, UniProt identifier, and any enzyme annotations were retrieved from KEGG [69]. GO term annotations were retrieved from UniProt [55]. If KEGG does not contain an enzyme annotation for the protein, we check if any of the GO terms correlate to an enzyme.
2) If the protein is not annotated with an enzyme, we assumed it to be a transporter. Otherwise, we assumed it to be an enzyme.
3) For every enzyme with sequence information in UniProt we created a Hidden Markov Model [185]. If three or more sequences of the enzyme were present in SwissProt, we based the model solely off of these proteins. In other cases, we based the model both off of SwissProt and Trembl proteins. All *C. curvatus* proteins were subject to a HMMScan [185] against any enzyme annotated to the *Y. lipolytica* protein.
4) For every *C. curvatus* protein we did a global alignment against the *Y. lipolytica* protein using NEEDLE [199].
5) We only took the top three *C. curvatus* candidates, starting with the most likely candidate until we found a match to the *Y. lipolytica* protein and its function. For enzymes the top candidates are selected using the HMMScan results; for transporters the top candidates are selected using the NEEDLE results.
6) The *C. curvatus* protein its GO term predictions were compared to the *Y. lipolytica* protein its GO terms. Information was returned on whether each *C. curvatus* GO term was identical or similar to a *Y. lipolytica* GO term, or not related to the *Y. lipolytica* protein at all.
7) We performed a web-BLAST for the *C. curvatus* protein on the SwissProt database to visually inspect if it has any significant homology, and if any of the homologous proteins have relevant functional information.
8) We performed a web-PFAM HMMScan for the *C. curvatus* protein to see if there are any known domains in the sequence, and if these domains have any function related to the *Y. lipolytica* protein.

9) To summarize, we have: a HMMScan on any enzyme of interest, a global alignment to the *Y. lipolytica* protein, comparison of GO terms between the proteins, homology information, and domain information. Using this information, we assess if the *C. curvatus* protein performs the same function as the *Y. lipolytica* protein. If not the case, we repeat the assessment steps with the next *C. curvatus* top hit.

For all *C. curvatus* and *Y. lipolytica* matches, we transferred the enzymatic or transporter function with its accompanying reaction to the *C. curvatus* protein.

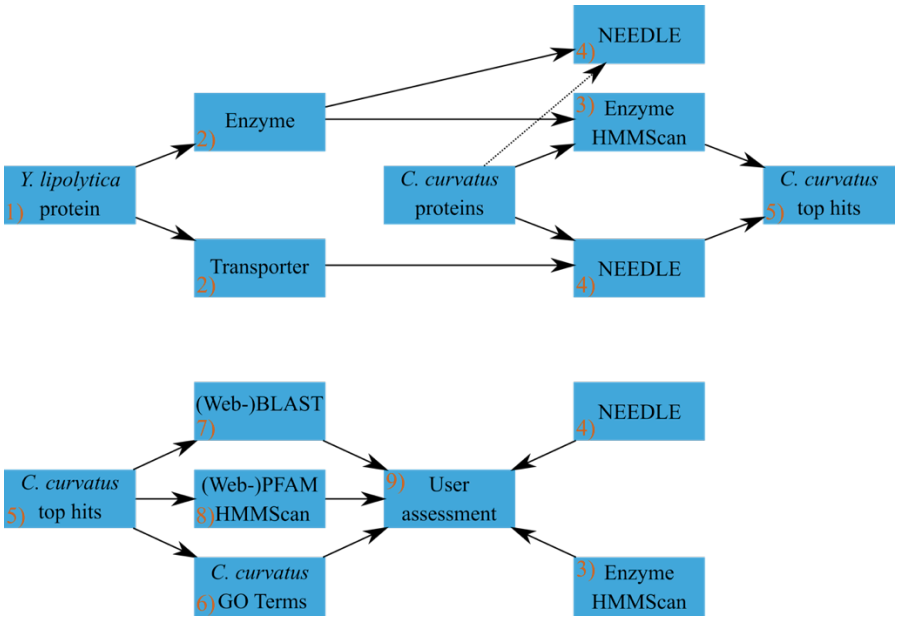


Figure 1: Workflow of the biocuration process. Numbered steps are further explained in the main text.

2.6 Differential expression analysis

We performed differential expression using the nitrogen rich and nitrogen starved RNA-Seq sets. This was done using the EDGE R package [200], with the replete set as a reference. Anything up-regulated with less than 0.05 p-value was taken as a protein up-regulated during nitrogen starved conditions; everything down-regulated with less than 0.05 p-value was taken as a protein down-regulated during nitrogen starved conditions. Triacylglycerol (TAG) metabolism analysis was done by taking all manually annotated proteins, color-coding the proteins based on their differential expression, and overlaying the proteins on KEGG maps [69].

2.7 GO enrichment analysis

Using the gene expression data, we performed a GO-enrichment analysis. This analysis was only done using predictions from CrowdGO, as the manually annotated GO terms are a specific subset of proteins. Initial GO enrichment was done using GOSEQ from the Bioconductor R package [201]. The resulting GO enrichments with their p-values were entered into REVIGO to produce more general GO term enrichments [202]. For this we used REVIGO's Lin's algorithm [188] and a similarity score of 0.4. The tables were taken directly from REVIGO, as were the figures apart from some minor human readability edits.

3 - Results and discussion

3.1 Gene prediction

We performed a de novo gene prediction on the genome of *Cutaneotrichosporon curvatus* ATC20509 [203] using BRAKER1.0. Translation of these genes results in some 7,600 proteins, of which around 7,400 are larger than 100 amino acids in length (Table 2).

Table 2: Protein summary of *Cutaneotrichosporon curvatus*

	Total
Proteins	7,597
Proteins larger than 33 AA	7,596
Proteins larger than 50 AA	7,593
Proteins larger than 100 AA	7,376
Average protein length	493
Largest protein	5.051

3.2.1 Cutaneotrichosporon curvatus annotation summary

All *C. curvatus* proteins were annotated using CrowdGO (**Chapter 4**). We found 168 proteins that did not have a BLAST hit to any protein in UniProt and treated these hypothetical proteins as false positive proteins. The CrowdGO annotations showed that a select amount of generic GO terms is annotated to these proteins (Supplementary table 1), and as such we assume these GO terms are prone to overfitting by CrowdGO. The full annotations are summarized in table 3 A without filtering for overfitted terms, and table 3 B with filtering for overfitted terms. A term was deemed overfitted if it appeared more than 40 times in the annotations for the hypothetical proteins in Supplementary table 1, and all the GO term annotations for any overfitted GO term were removed from the list.

3.2.2 *Cutaneotrichosporon curvatus* annotation summary - discussion

What can be observed in the annotation summary is that merging the Argot2, FFPred2, and InterProScan predictions with CrowdGO results in lower amount of GO terms than the sum of the three tools after filtering for overfitted terms. This is contradictory with the results from **chapter 4 table 3** where it shows that CrowdGO re-annotates a sizeable amount of GO terms initially labelled as false negative to a true positive, and true negative to a false positive. However, the test set proteins of chapter 4 did not contain hypothetical proteins, which affects the results.

The decrease of total annotated GO terms, and average GO terms per protein, is a clear indicator CrowdGO performs strict filtering of the input GO terms.

A noticeable result is the difference in the prediction of Cellular Component GO terms before and after filtering out of overfitted terms. This is due to two factors: only two of the input tools are able to predict Cellular Component terms, and because many Cellular Component GO terms are relatively non-specific compared to Biological Process or Molecular Function terms but still technically correct predictions for many terms, CrowdGO is prone to overfitting on the data with respect to these terms. This was noticeable in the annotation of the presumed false positive proteins, where over half of the over-fitted terms were Cellular Components explaining the dramatic reduction in proteins annotated with a Cellular Component term by CrowdGO (table 3).

Table 3: Annotation numbers for CrowdGO and the three methods utilized in the CrowdGO. Displayed are the amount of proteins having one or more GO terms annotated to it, the range of GO terms used to annotate these proteins, the total amount of GO terms annotated to proteins, and the average amount of GO terms annotated to a protein. MF: Molecular Function. BP: Biological Process. CC: Cellular Component.

A: without filtering out overfitted terms.

Method	Annotated proteins MF	Annotated proteins BP	Annotated proteins CC	Unique GO's	Total GO's	GO's per protein
CrowdGO	5,809	6,621	7,566	988	21,566	3
Argot2	2,962	3,032	1,441	878	4,742	2
FFPred2	4,462	5,076	0	175	11,887	2
InterProScan	3,086	2,655	1,293	574	4,104	1

B: After filtering out overfitted terms

Method	Annotated proteins MF	Annotated Proteins BP	Annotated Proteins CC	Unique GO's	Total GO's	GO's per protein
CrowdGO	3,091	2,764	473	988	21,566	3
Argot2	2,962	3,032	1,441	878	4,742	2
FFPred2	4,405	5,076	0	175	11,887	2
InterProScan	3,064	2,655	998	574	4,104	1

3.3.1 Annotation comparison to *Cutaneotrichosporon oleaginosum*

We compared the *C. curvatus* predictions to the existing UniProt annotations of *Cutaneotrichosporon oleaginosum*, both after filtering for over-fitted terms, to provide a reference on the quantitative performance of CrowdGO (Table 4A). Additionally, we compared the annotations between the species for the 5,000 cross-species orthologs found (Table 4B).

3.3.2 Annotation comparison to Cutaneotrichosporon oleaginosum discussion

The comparison between the CrowdGO annotations of *C. curvatus* and the UniRule annotations of *C. oleaginosum* shows that they annotate roughly the same amount of proteins. Importantly, the range of GO terms used by CrowdGO and UniRule is vastly different. In *C. curvatus* the 5,000 orthologs are annotated with only 839 GO terms, compared to 1,898 for UniRule (Table 4 B). It is certainly possible that UniRule is able to correctly annotate a wider range of GO terms to proteins, however over 1,800 compared to less than 850 for CrowdGO indicates that at least a fraction of these GO terms are false positives. Additionally, using the predictions of more than three tools as an input for CrowdGO will likely increase the range of GO terms it uses to annotate proteins, but this will likely also increase the number of false positives.

Finally, table 4 B shows that approximately two thirds of orthologs between the two species have one or more GO terms in common. This implies that both the CrowdGO and UniRule annotations are able to represent these proteins, and that the range of GO terms assigned by CrowdGO to *C. curvatus* is large enough to represent the putative function of its proteins.

Table 4: Comparison of the *Cutaneotrichosporon curvatus* annotations with the *Cutaneotrichosporon oleaginosum* protein annotations.

A: Comparison of all proteins

	Cutaneotrichosporon curvatus (CrowdGO)	*Cutaneotrichosporon oleaginosum (UniRule)*
Total proteins	7,597	8,317
Proteins annotated	4,332	4,223
Total GO terms	8,123	9,569
Unique GO terms	952	2,261
GO's per protein	1,9	2,3

B: Comparison of orthologs

	Cutaneotrichosporon curvatus (CrowdGO)	*Cutaneotrichosporon oleaginosum (UniRule)*
Total proteins	5,110	5,110
Proteins annotated	2,946	2,914
Total GO terms	5,344	6,828
Unique GO terms	839	1,898
GO's per protein	1,8	3,6
Matching ortholog annotations	2,342	2,342

3.4 Manual annotation of the *Cutaneotrichosporon curvatus* metabolic proteins

We manually curated the metabolic proteins of *C. curvatus* using the CrowdGO annotations and comparative genomics with proteins from the *Yarrowia lipolytica* model iNL895[198] (Methods section 2.5). This resulted in 710 manually annotated proteins, involved in over a thousand reactions (Table 5). Additionally, we compared the CrowdGO annotations to the manual curations of *C. curvatus* in table 6. In this table, closely related annotations are proteins that have one or more GO parent and child terms of each other. During the manual annotation process the CrowdGO annotations of 553 proteins were in line with the final manual annotation, according to the biocurator.

Table 5: Summary of the manual annotations

	Total
Proteins	710
Enzyme proteins	540
Transporter proteins	170
Enzyme annotations	930
Unique enzyme annotations	461
Reactions	1155

Table 6: Annotation comparison between the high-throughput and the manually annotated proteins. Only enzymatic proteins were chosen for analysis.

Type	Amount
Exact annotation	110
Closely related annotation	293
False positive annotation	74
Annotation used in man annotation	553

3.5 Manual annotation of the Cutaneotrichosporon curvatus metabolic proteins discussion

In our process of manually annotating the C. curvatus proteins we used our CrowdGO annotations as part of the curation pipeline. Comparing the CrowdGO numbers to the manual annotations provide valuable insight in the usefulness of the predictions for biological interpretation (Table 6). With a 0.6 cut-off for the predicted annotations, a minority fraction of the proteins has an exact match with its manual annotated enzymatic function. When looking at GO terms that are related by either being a direct or indirect term, a majority fraction of the predictions match with the manual annotations. Only 74 proteins are annotated without an exact or closely related annotation to the proteins biocurated function, which we deemed false positives. This is after removing the term GO:0016887 ATPase activity from all our predicted annotations, CrowdGO annotated it 44 times above and never below the 0.6 confidence interval, indicating over-fitting by CrowdGO. Not removing ATPase activity from the set of annotations would increase the amount of false positive annotations to over a hundred.

3.6 GO Enrichment analysis

We use the CrowdGO annotations and differential expression analysis as proxy for protein activity to obtain an overview of *C. curvatus* triacylglycerol (TAG) synthesis in nitrogen replete and deplete conditions. GO enrichment analyses for these conditions are summarized for replete conditions (Figure 2,3) and nitrogen deplete conditions (Figure 4,5). Replete conditions show more expression for cell maintenance and growth-related proteins, while nitrogen starved conditions show more expression for proteins related to catabolism and usage of stored energy.

3.6.1 GO Enrichment analysis discussion

The GO enrichment analysis does not provide much more information than what is already known for oleaginous eukaryotes in relation to nitrogen stress conditions. However, because the general processes that are differentially regulated are so well known, for example cell growth during normal conditions and catabolism during nitrogen stress conditions, we can use this as a validation of the high-throughput CrowdGO annotations on a general level. Considering the GO enrichment shows us what we would expect, we assume that the CrowdGO annotations are accurate on a general-level basis.

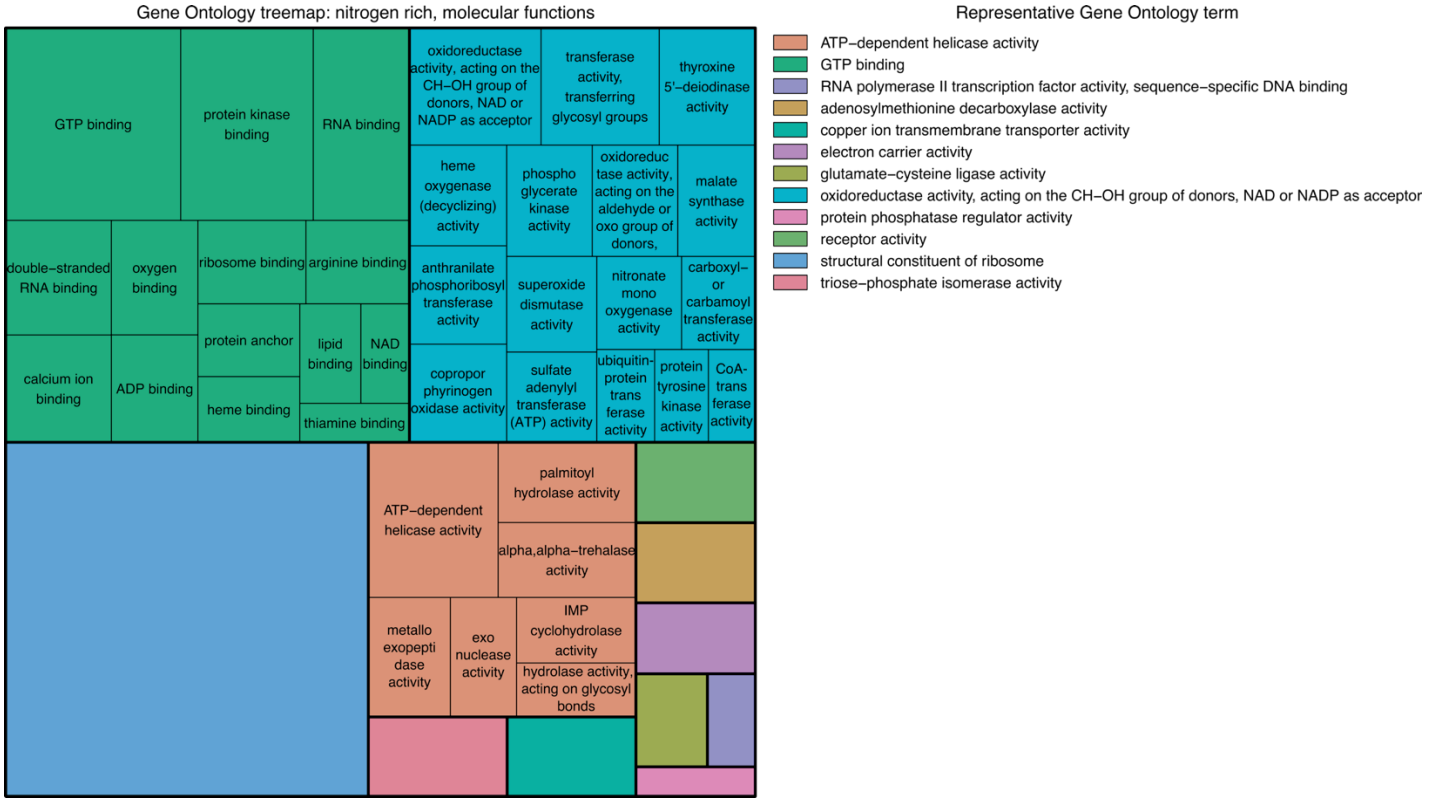


Figure 2: Gene ontology enrichment of *C. curvatus* for nitrogen rich conditions and molecular functions. Colors represent GO term classes, box sizes are relative to the GO's p-value for up-regulation under the growth conditions.

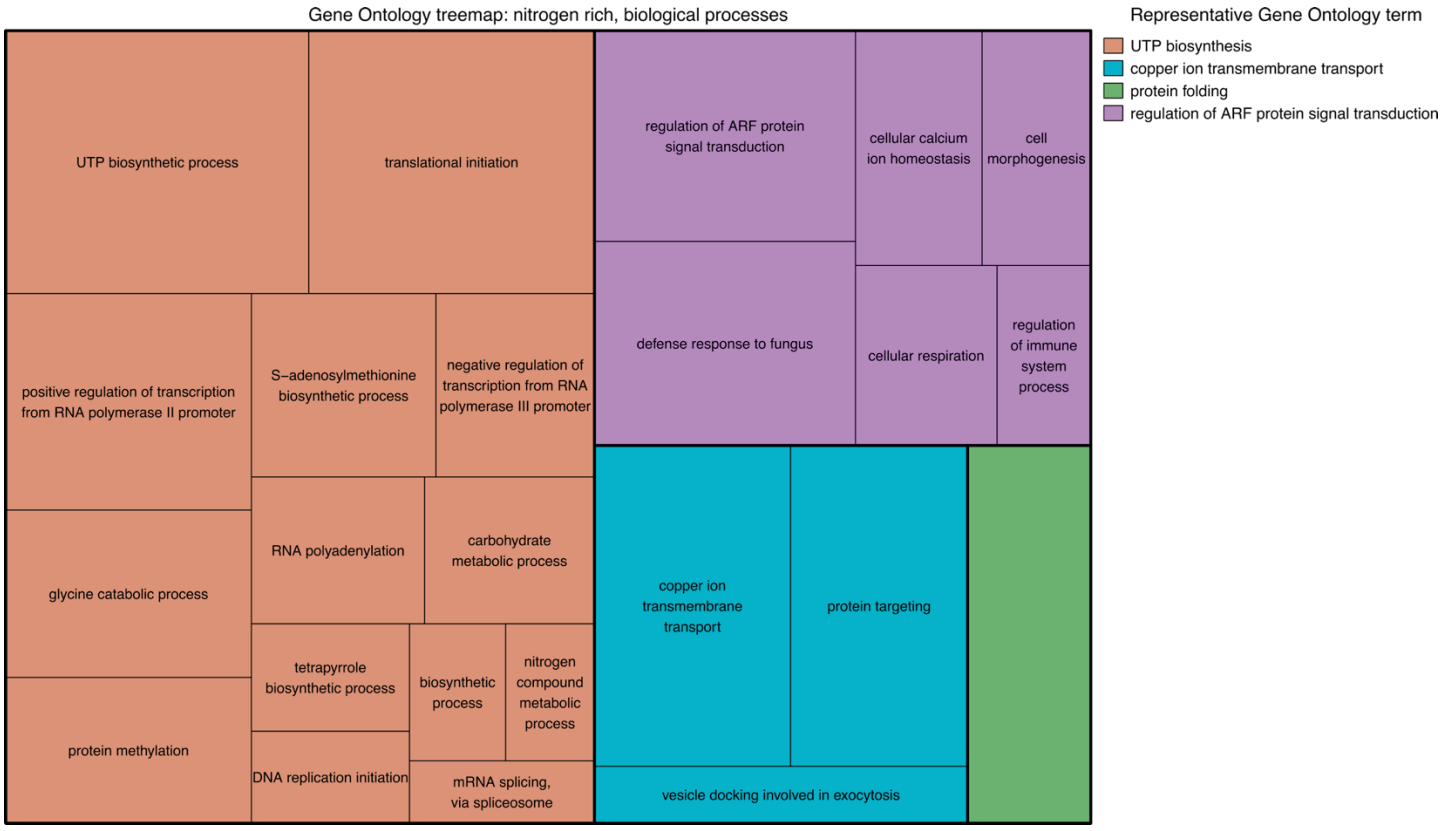


Figure 3: Gene ontology enrichment of *C. curvatus* for nitrogen rich conditions and biological processes. Colors represent GO term classes, box sizes are relative to the GO's p-value for up-regulation under the growth conditions.

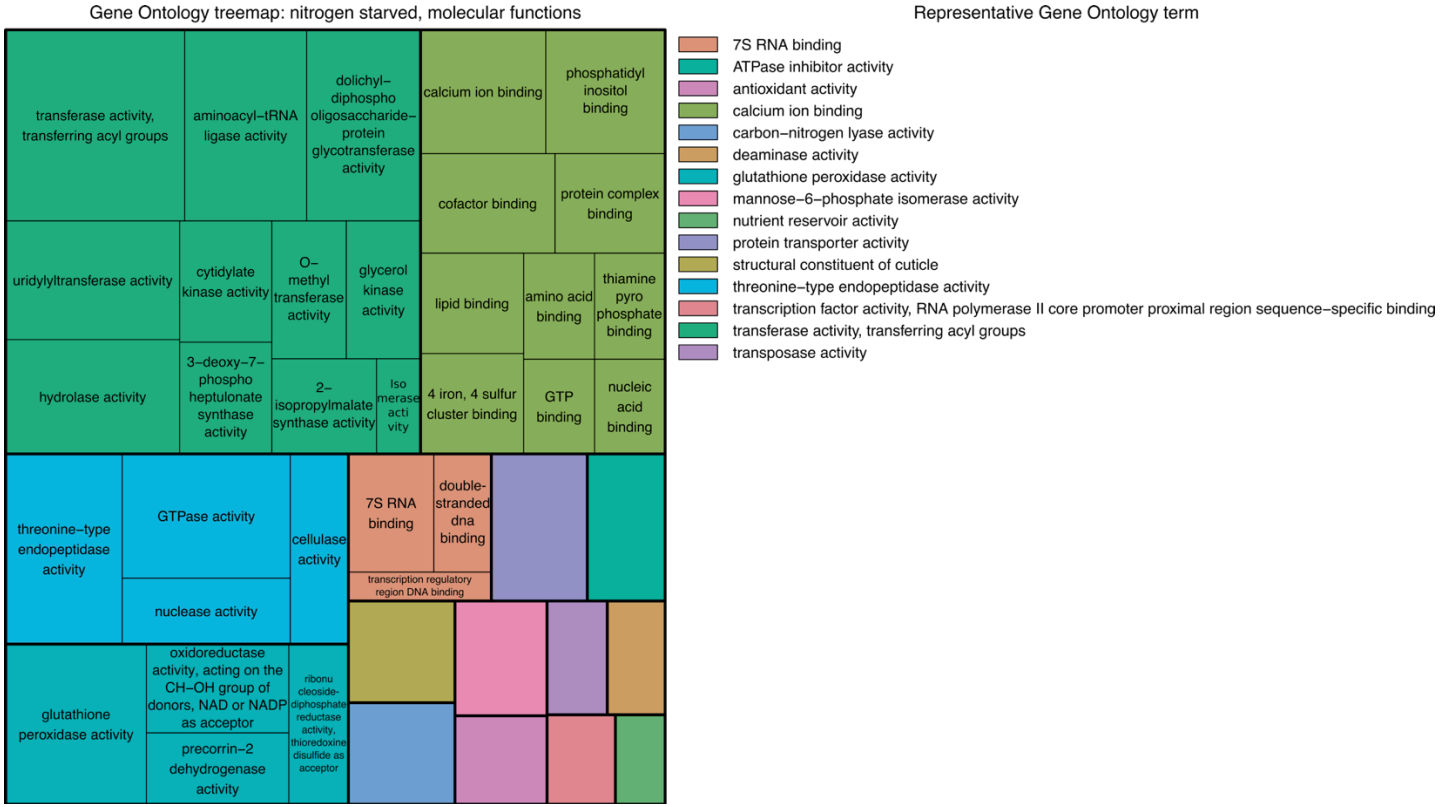


Figure 4: Gene ontology enrichment of *C. curvatus* for nitrogen starved conditions and biological processes. Colors represent GO term classes, box sizes are relative to the GO's p-value for up-regulation under the growth conditions.

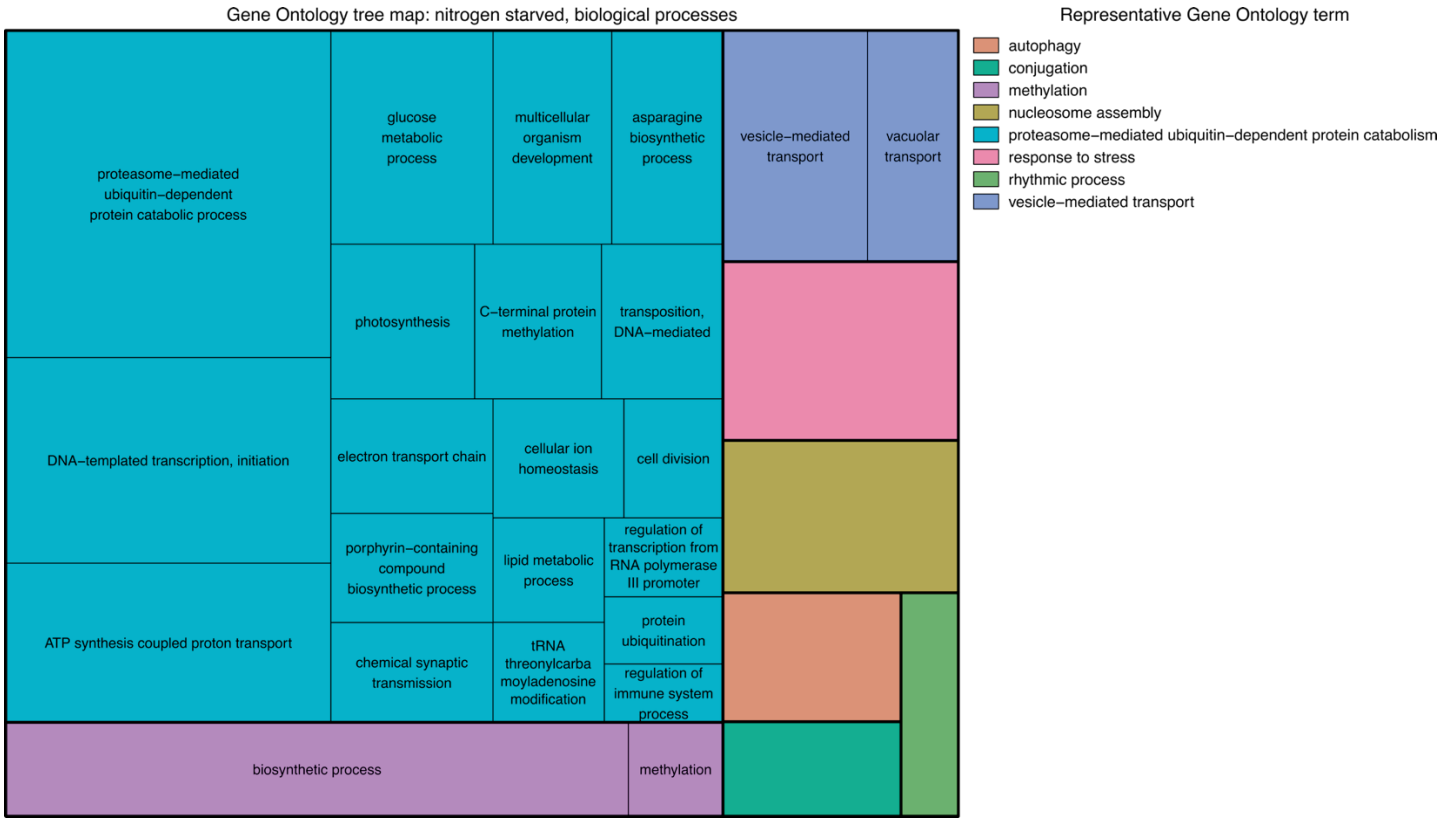


Figure 5: Gene ontology enrichment of *C. curvatus* for nitrogen starved conditions and molecular functions. Colors represent GO term classes, box sizes are relative to the GO's p-value for up-regulation under the growth conditions.

3.8 Differential expression analysis of triacylglycerol synthesis

We use our manual annotations and differential expression to characterize TAG synthesis in *C. curvatus* (Figure 6). Several key enzymes show a significant ($p<0.05$) up-or-down-regulation during nitrogen starved conditions. These enzymes are down-regulating sugar metabolism, differentially expressing the TCA cycle, and differentially expressing various steps between glycerol and triacylglycerol conversion. EC 3.1.1.23 acyl glycerol lipase, which is present for *Y. lipolytica* in the KEGG database [69], was missing in the *Y. lipolytica* genome-scale model. A candidate for the enzyme was found in *C. curvatus* and was manually annotated without comparative genomics. The fold changes for the enzymes can be found in Table 7.

3.8.1 Differential expression analysis of triacylglycerol synthesis - discussion

The differential expression analysis shows us a complete picture of *C. curvatus* TAG synthesis, with only one key enzyme missing in the entire set of 33 enzymes used to annotate the differential expression of TAG metabolism. Differential expression reveals up-and-down regulation of key processes, such as down-regulation of glucose and fructose metabolism related enzymes during nitrogen depleted conditions, confirm our annotations. Additionally, we see other interesting enzymes in response to nitrogen-depleted conditions, such as an up-regulation of enzyme 2.3.1.158 phospholipid diacylglycerol acyltransferase (PDAT) compared to enzyme 2.3.1.20 diglyceride acyltransferase (DGAT). PDAT is shown in literature to be a big contributing enzyme for oleaginous yeasts and adding additional copies of this enzyme leads to increased TAG synthesis. Another interesting finding is the down-regulation of enzyme 2.7.1.30, which is responsible for the direct conversion of glycerol to glycerol-3P. The fact that during nitrogen starved conditions *C. curvatus* is suggested to create glycerol-3P through glycerone-phosphate might be interesting for any follow up studies regarding TAG synthesis in this organism.

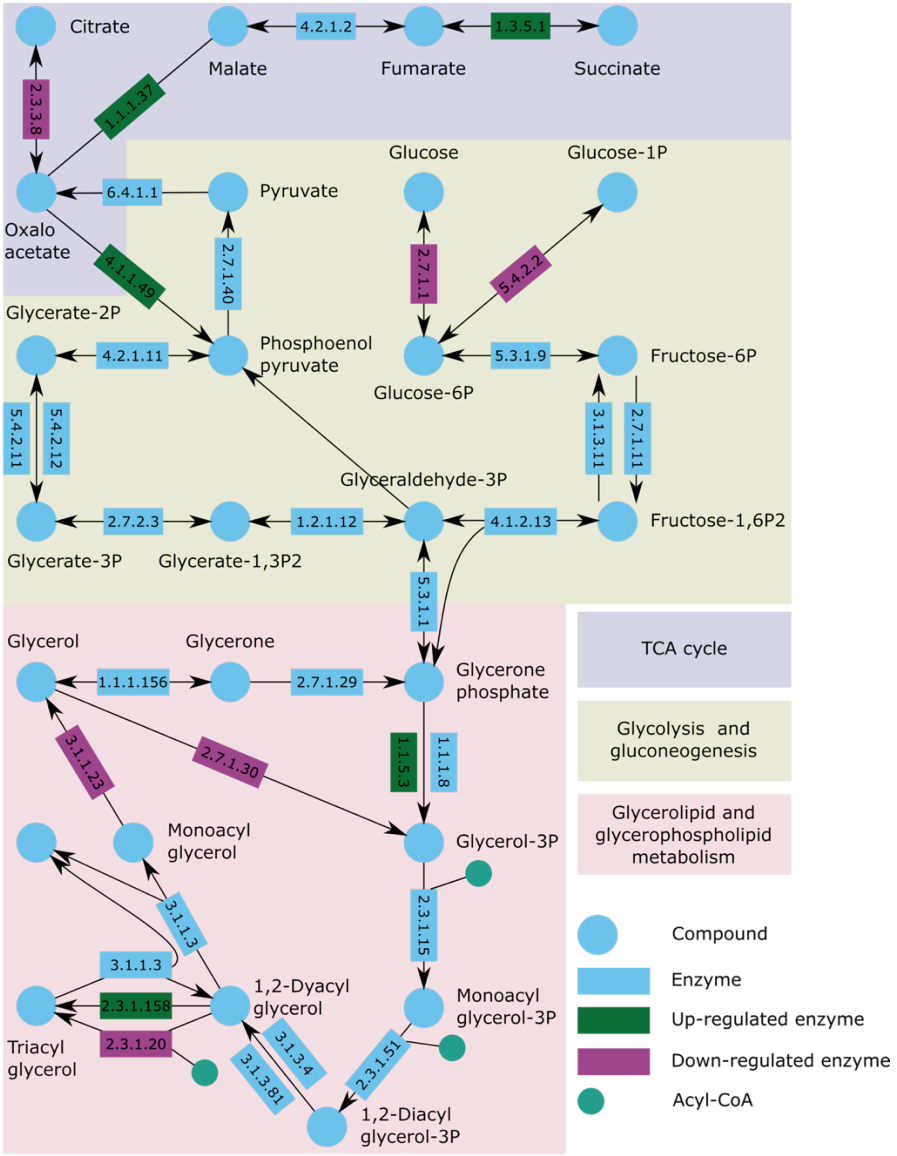


Figure 6: Summary of TAG synthesis and key differentially regulated enzymes in *Cutaneotrichosporon curvatus* during nitrogen deplete conditions.

Table 7: LOG2 fold changes of the differentially expressed enzymes related to triacylglycerol synthesis, as displayed in figure 6

Enzyme	Fold change	Adjusted P-value
2.3.3.8	-0.32	0.46
2.7.1.1	-0.21	0.006
5.4.2.2	-0.20	0.01
3.1.1.23	-0.33	0.16
2.7.1.30	-0.56	1.58e-10
2.3.1.20	-0.31	0.003
1.3.5.1	0.28	0.01
1.1.1.37	0.43	0.04
4.1.1.49	0.48	5.6e-5
1.1.5.3	0.26	0.004
2.3.1.158	0.24	0.005

Conclusion

We were able to thoroughly functionally annotate *C. curvatus* proteins using the CrowdGO method as described in chapter 4, as shown by comparisons with *C. oleaginosum* existing annotations, GO enrichment analysis, and its role in manually annotating the proteins. However, the comparison to the manual annotations show that while the CrowdGO method is a clear improvement over existing methods, it is not able to consistently predict full enzyme annotations. That said, any analysis of specific metabolic processes should be done by manual annotation. The CrowdGO annotations have been specifically useful in speeding up this process.

Chapter 6

Comparing the same reveals the difference: systems biology of *Botryococcus braunii* Races A and B

SPLASH WP2 Consortium: Douwe van der Veen[1*], Eugen Urzica[2*], Maarten Reijnders[3*], Carolyn M.C. Lam[4], Sven Warris[5], Olga Blifernez-Klassen[6], Joao D. Gouveia[1], Swapnil Sudhakar Chaudhari[6], Johannes Leufken[2], Doris Gangl[7], Mark A. Scaife[7], Henri van de Geest[5], Linda Bakker[5], Jörn Kalinowski[6], Jan Springer[1], Olaf Kruse[6], Vitor Martins Dos Santos[3,4], Peter Schaap[3], Sander A. Peters[5], Alison G. Smith[7], Michael Hippler[2]

1. Bioprocess Engineering Group and AlgaePARC, Wageningen University and Research, P.O.Box 16, 6700 AA Wageningen, The Netherlands
2. Institute of Plant Biology and Biotechnology, University of Münster, Münster 48143, Germany
3. Laboratory of Systems and Synthetic Biology, Wageningen University, Stippeneng 4, Building 124 (Helix), 6708 WE Wageningen, The Netherlands.
4. LifeGlimmer GmbH, Markelstrasse 38, 12163 Berlin, Germany.
5. Business Unit of Bioscience, Cluster Applied Bioinformatics, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands.
6. Bielefeld University, Faculty of Biology, Center for Biotechnology (CeBiTec), Universitätsstrasse 27, 33615 Bielefeld, Germany
7. Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK

*These authors contributed equally

Manuscript in preparation

Abstract

Motivation: *Botryococcus braunii* is a biotechnologically interesting microalgae to study due to its ability to synthesize and excrete high amounts of hydrocarbons or polysaccharides, depending on the strain. In this chapter we perform a comparative genomics study to find differences between these two strains, which we can correlate to the production of hydrocarbons or exopolysaccharides. However, *Botryococcus braunii* is a non-axenic microalga, which is challenging to perform a genomics study on. Therefore, we also set out to create a workflow for future microalgal research, able to study non-axenic microalgae both on a genome-scale and a pathway scale.

Results: We developed a proteomics-based workflow, which with the help of CrowdGO (chapter 4) and manual annotations provided insights in the comparative genomics of *Botryococcus braunii*. Using this workflow, we identified several key enzymes involved in hydrocarbon synthesis and exopolysaccharide synthesis. In the non-mevalonate pathway, we identified 4 enzymes that are significantly more expressed in the hydrocarbon producing strain, and in the GDP-L-Fucose biosynthesis pathway we identified 3 enzymes that are significantly more expressed in the polysaccharide producing strain. We also found significant differences in expression levels of key enzymes involved in the core metabolism of *Botryococcus braunii*, showing that hydrocarbon and exopolysaccharide production have a metabolism-wide effect on the species, and indicating the two strains might not be as closely related as their shared name suggests.

1. Introduction

Botryococcus braunii, a member of the class *Chlorophyceae*, is a prime example of a photosynthetic microalgae exploitable for commercial interest, due to its unique ability to produce and excrete vast quantities of large chain hydrocarbons and exopolysaccharides during nitrogen limiting conditions [204]. The hydrocarbons form a lipid biofilm matrix holding together *Botryococcus braunii* cells in colonies, and the exopolysaccharides are believed to serve as carbon reserves and perhaps as a protective layer against harmful environmental conditions and pathogens. These features attract attention as a potential resource of natural biopolymers, aromatic bulk chemicals, and fine chemicals.

The *Botryococcus braunii* species are classified into three chemical races: A, B, and L, depending on the chemical nature of the large chain hydrocarbons they

synthesize [205, 206]. Race A primarily synthesizes odd-numbered n-alkadiene and n-alkatrienes (C23-C33); race B primarily synthesizes triterpenoid botryococcenes (C30-C37) and methylated squalenes (C31-C34); race L primarily synthesizes the tetraterpenoid hydrocarbon lycopadiene, via an alternative pathway using a squalene synthase-like enzyme [207]. The predominant metabolic pathway for the biosynthesis of isoprenoids is the non-mevalonate (MEP) pathway [208]. In this pathway, 1-deoxy-D-xylulose-5-phosphate (DOXP) is converted into 2-C-methyl-D-erythritol-4-phosphate and MEP intermediate, which in turn is converted into isopentenyl diphosphate (IPP) and dimethylallyl (DMAPP), which are used as precursors for the biosynthesis of farnesyl pyrophosphate (FPP). FPP is subsequently used via a series of intermediates for the synthesis of botryococcenes and methylated squalenes. Carotenoid biosynthesis follows the same pathway as botryococcenes, but starting from the higher carbon number geranyl-geranyl pyrophosphate [204]. However, more evidence is necessary to fully understand the characteristic metabolic pathways underlying large chain hydrocarbon and carbohydrate biosynthesis in *Botryococcus braunii*.

Botryococcus braunii is, however, a challenging species to study. The genus has an unclear taxonomy, and its species boundaries are not well defined. Phenotypic characteristics such as colony forms, colour, cell shape, and cell size, have been reported to depend on environmental growth conditions hampering its diagnostic use [209-211]. In addition to morphological data, molecular data have been used to increase the phylogenetic resolution. Kawachi *et al* defined relationships based on nuclear 19S rDNA and categorized 31 isolates into three major phylogenetic clades correlating to a high degree with the chemical races A, B, and L [212]. Recently, Hegedüs *et al* used 18S rDNA and ITS2 molecular markers to classify *Botryococcus* race A strains and defined two distinct phylogenetic subclades: A1 and A2 [213]. The phylogenetic relationship of *Botryococcus* strains appeared to correlate with the typical hydrocarbon profile of race A, B, and L. However, the high genetic divergence of *Botryococcus braunii* does not support their classification into one single species. This high genetic divergence is reflected in the transcriptome, which hints at substantial differences in the biosynthetic pathways that *Botryococcus* strains have adopted for the synthesis of hydrocarbons, ether-lipids, and polysaccharides. In addition, several studies have indicated *Botryococcus* is difficult to maintain and grow as an axenic culture, complicating transcriptome profiling [116]. Specifically, *Rhizobium* species have been

reported to encourage the growth of *Botryococcus*, while *Acinetobacter* species have been reported to have a negative interaction [128], further complicating the metabolic profiling *Botryococcus* .

We have undertaken a multi-disciplinary approach to identify common features in non-axenic *Botryococcus braunii* CCALA778 and AC761 strains, which previously have been classified as race A and race B strains respectively [214, 215]. We provide a comparative overview that links changes in algal transcripts and proteins to the abundance of characteristic metabolites in race A and B non-axenic cultures, and have identified candidate targets that play an essential role in the characteristic metabolism of *Botryococcus*.

2. Materials and Methods

2.1 Organisms and their cultivation

Botryococcus braunii strains CCALA778 and AC761 were obtained from the Culture Collection of Autotrophic Organisms (Trebon, Czech Republic) and Algobank Caen (Caen, France), respectively. Both strains were maintained in 250 mL Erlenmeyer flasks in an Infors HT Multitron incubator with the following environment parameters: illumination: Philips FL-Tube L 36W/77 lamps with intensity set at 150 umol photon m-2 sec-1; light:dark photoperiod 18:6 h; 2.5 % percent CO2; temperature 25°C; mechanical shaking at 90 rpm.

Culture media consisted of modified Chu 13 medium [216] without citric acid, with the following composition: 1200 mg L-1 KNO3, 200 mg L-1 MgSO4.2H20, 108 mg L-1 CaCl2.2H2O, 104.8 mg L-1 K2HPO4, 20 mg L-1 Fe-Na2EDTA, 9.4 µg L-1 Na2O4Se, 2.86 mg L-1 H3BO3, 1.8 mg L-1 MnSO4.4H2O, 220 µg L-1 ZnSO4.7H2O, 90 µg L-1 CoSo4.7H2O, 80 µg L-1 CuSO4.5H2O, 60 µg L-1 Na2MoO4.2H2O, 10 µl L-1 H2SO4. Final pH was adjusted to pH 7.2 with NaOH.

2.2 Photobioreactor cultivation

For experiments, the strains were cultivated in flat-panel airlift photobioreactors (Algaemist, Technical Development Studio, Wageningen University and Research, The Netherlands) with a working volume of 0.4 L, and optical depth of 14 mm, and an illuminated area of 0.028 m2. All runs were operated in batch mode. Light was provided by LED lamps (BXRA W1200, Bridgelux, USA) with a warm-white spectrum. The photon flux density was measured with a LI-COR 190-SA 2pi PAR (400-700 nm) quantum sensor (LI-COR, USA). Incident light intensity was measured at 28 points evenly

distributed over the light-exposed surface of the front glass panel of the culture chamber, and light measurement was repeated for every experiment. The applied light regime was a block 18h:6h day:night light:dark cycle, and the incident light intensity set at averaging 150 μmol photon m^{-2} sec^{-1}.

Aeration and mixing were done by sparging a gas stream of filtered air at a rate of 200 mL min^{-1} (0.5 vvm). pH was maintained at 7.2 (PLUS MINUS 0.1) by on-demand addition of CO_2 to the airflow. To ensure sufficient CO_2 in the medium $NaHCO_3$ was added to the medium to a final concentration of 5mM. Temperature was set to 25°C. The bioreactors were inoculated with shakeflask-derived biomass so that the initial OD 680nm = 0.2; addition of biomass was considered the start of the experiment and is referred to as t = 0.

2.3 Experimental design

Earlier bioreactor tests with both strains indicated that, with using an initial KNO3 concentration of 1200 mg L^{-1} and using the initial inoculation density, a window of non nutrient-limited growth of between 8 to 12 days was achievable, counted from the first day where biomass concentration reached 1 g L^{-1}. The end of this period was marked by the lack of detecting nitrate in the medium and the reporting of the steady low signal reported by the bioreactor's light-out sensor.

Per strain, at minimum 5 bioreactor runs, using inoculum obtained from independent shake-flask cultures, were operated. Two of these runs were sacrificed when nitrate was depleted; three runs continued for 7 days after nitrate was depleted. To facilitate quantitative proteomics analysis, per run either ^{15}N-labeled KNO3 or ^{14}N KNO3 (Sigma-Aldrich) were used.

Prior to every sampling time point, 3 mL culture broth was removed from the bioreactor to clear the sampling port. Samples to determine cell dry weight (5 mL), optical density (1 mL). chlorophyll content (1 mL), and nitrate content (1 mL) were taken daily; samples for carbohydrate or hydrocarbon analysis (1 mL) or elemental analysis (10 mL) were taken intermittently, as omics sampling had precedence.

To acquire sufficient biomass for the various omics technologies, sampling took place on alternating days to limit large disturbances in the cultivation environment because of the large sampling volumes; total sampling volume was kept as low as possible and did not surpassed 15% of bioreactor working volume.

For metabolite analysis and transcript expression analysis, 10 mL of culture broth were removed, while for proteomics analysis, 20 mL culture broth was removed. Samples were added to prepared tubes containing equal volume of ice-cold (-20 °C) methanol, spun for 2 min at 4,000 x g in a bench-top centrifuge, and supernatant was decanted while taking care to not disturb the pellet. Samples were snap-frozen in liquid nitrogen and stored at -80 °C until analysis. All omics samples were snap-frozen within 4 minutes after reactor sampling.

After sampling, medium without nitrate was added to the reactor vessel up to the working volume. When nitrate was present in the medium, nitrate was added to the bioreactor to the initial KNO_3 concentration, thus ensuring that removal of large sampling volumes did not remove nitrate dissolved in the medium.

2.4 Cell Dry Weight

Five mL aliquots of culture broth were filtered onto pre-weighed GF/D glass-fibre membranes (Whatman) and washed with 5 mL demineralized water. The filters were dried at 100°C for 24 hours and weighted, after which the biomass amount was determined by subtraction. For biomass concentrations over 3 g L^{-1}, filters became clogged and 2 mL aliquots of culture broth were filtered instead.

2.5 Hydrocarbon extraction and analysis

One mL of culture broth was combined in a glass vial with 2.5 mL methanol and 1.25 mL dichloromethane and mixed on a rotary shaker at 30 rpm for 6 hours. After mixing, 1.25 mL dichloromethane was added and mixed for one minute, after which 1.25 mL 0.9 % (w/v) NaCl solution was added and mixed for another minute. Hereafter, samples were centrifuged for 5 minutes at 1,500 x g. The bottom phase was removed to a new glass vial using a glass Pasteur pipette and dried under nitrogen gas. The residue was resuspended in 1.0 mL hexane and stored at -20°C.

Hydrocarbon analysis was carried out using gas chromatography (GC-FID). The instrument used was an Agilent Technologies HP6890 series equipped with auto sampler, a using Restek Rxi-5ms (30 m x 0.25 mm x 0.25 µm) column. Helium was used as the carrier gas, and a hydrogen/air moisture detection, gas splitless injectors at 350°C oven temperature and injection volume of 1 µL. The oven program was 50°C for 1 minute, then 15°C per minute to 180°C, then 7°C

per minute to 230°C, then 30°C per minute to 350°C and hold for 15 minutes with a total running time of 35 minutes. Samples were diluted in hexane, and several dilutions of standards using squalene were used.

2.6 Total Carbohydrate extraction and analysis

Total carbohydrate contents were determined using the method first reported by Dubois [217]. In brief, 500 µL -20 °C methanol was added 500 µL culture broth, centrifuged for 3 minutes at 3,000 x g to pellet cells, and supernatant was carefully discarded. The pellet was hydrolysed by adding 500 µL of 2.5M HCl and incubation for 3 h while vortexing hourly and neutralized thereafter by adding 500 µL of 2.5M NaOH solution. Samples which 450 µL demineralized water was added. Gently, 500µL 5% phenol in water solution was pipetted into the tube. 2.5 mL concentrated sulphuric acid was added directly onto the liquid surface, and incubator at room temperature for 10 minutes. Hereafter, tubes were placed in a 35 °C waterbath for 30 minutes, while vortexing every 5 minutes. Absorbance was read at 483 nm. A D-glucose solution was used as standard.

2.7 Nitrate content

Measurements were performed with 110020 MQuant Nitrate Test according to manufactures protocol. The nitrate concentration is measured semi-quantitatively by visual comparison of the reaction zone of the test strip with the fields of a colour scale.

2.8 Phylogeny

All chlorophyta chloroplast sequences and an *Arabidopsis thaliana* chloroplast sequence were retrieved using the sequences from Lemieux et al [218]. We performed gene prediction on these chloroplasts using Prodigal [219]. For all these genes we did a Needle [199] search against 45 *Chlamydomonas reinhardtii* chloroplast proteins from the SwissProt database [55]. All the best Needle hits for each chloroplast sequence were extracted. The best Needle hits for each chloroplast protein were aligned using Muscle [220], and poorly aligned regions were removed using trimAl [221]. Finally, the alignments for all the chloroplast proteins were concatenated using PhyUtility [222]. The resulting concatenated alignment file was used as an input to PhyML to calculate the phylogenetic distances [223], and the final tree was drawn with TreeDyn [224].

2.9 Functional protein annotation

We assigned Gene Ontology (GO) terms using a combination of protein function classification methods: Argot2 [2], FFPred2 [167], and InterProScan [61].

Only annotations with a confidence of 0.7 or higher were annotated to the proteins.

For the pre-trained random-forest model we used the sequences of all proteins created between 01-11-2013 and 9-11-2015 with one or more experimentally validated GO terms. The number of true positives and false positives in the random-forest training was reduced to equal size in order to avoid bias in the dataset.

3. Results

3.1 Strain selection and cultivation

We selected two out of 16 *Botryococcus braunii* strains based on their race type and distinct product profile [215] as we aimed to contrast their underlying physiologies. Under our growth conditions the race A strain CCALA778 produces polysaccharides up to 2 g/L, while hydrocarbons could not be detected (Figure 1A). In contrast, the race B strain AC761 produces both polysaccharides and hydrocarbons although the former to a much lower extent (0.6 g/L) compared to CCALA778 (data not shown). Hydrocarbon levels in AC761 reached up to 0.4 g/L (Figure 1C). Colonial morphology differed between the strains (Figure 1B and D). CCALA778 cells appear round and assemble into compact colonies while AC761 cells tend to be droplet-shaped within a more dispersed colony.

Both races reached similar dry weights of 4.6g/L for CCALA778 and 5.2 g/L for AC761 (Figure 1A and C). Nitrogen levels were monitored in the bioreactors over the entire growth period. Increased product accumulation correlated with nitrogen-limited conditions in both races.

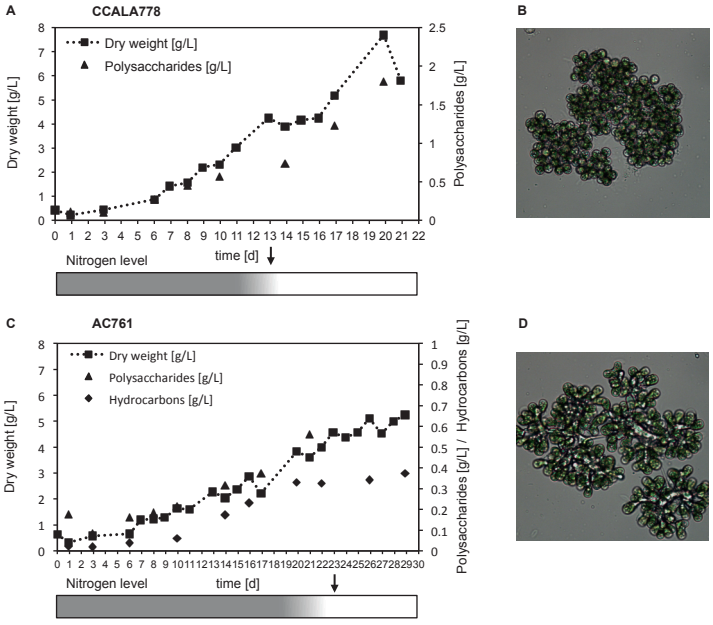


Figure 1: A) Hydrocarbon and polysaccharide production for the race A strain CCALA778, and B) its morphology. C) Hydrocarbon and polysaccharide production for the race B strain AC761 and D) its morphology.

3.2 Phylogenetic placement of *Botryococcus braunii*

The heterogeneity of chemical and morphological features has been recognized and it has been suggested that the name *B. braunii* in fact covers multiple distinct species [209], but the community has not widely embraced this nomenclature. Kawachi et al constructed a relationship between hydrocarbons produced and molecular phylogeny by 18S rRNA mapping between strains and showed that the chemical races map to distinct phylogenetic clades [212]. However, the relative positioning of *B. braunii* strains relative to other algae has not been carried out, in part due to the lack of sequencing information. Using our sequence information (see below), we constructed a phylogenetic distance tree of algal chloroplast proteins (Figure 2). Our analysis indicates that our two strains used lie too far away to be considered near-identical species, even though AC761 and CCALA778 are closer related to each other than to other microalgae. Reassuringly, CCALA778 is more closely related to the race A SHOWA strain, than it is to the race B AC761.

Following these results, we performed extensive comparative omics analysis between *Botryococcus braunii* CCALA778 and AC761. This workflow is visualized in figure 3.

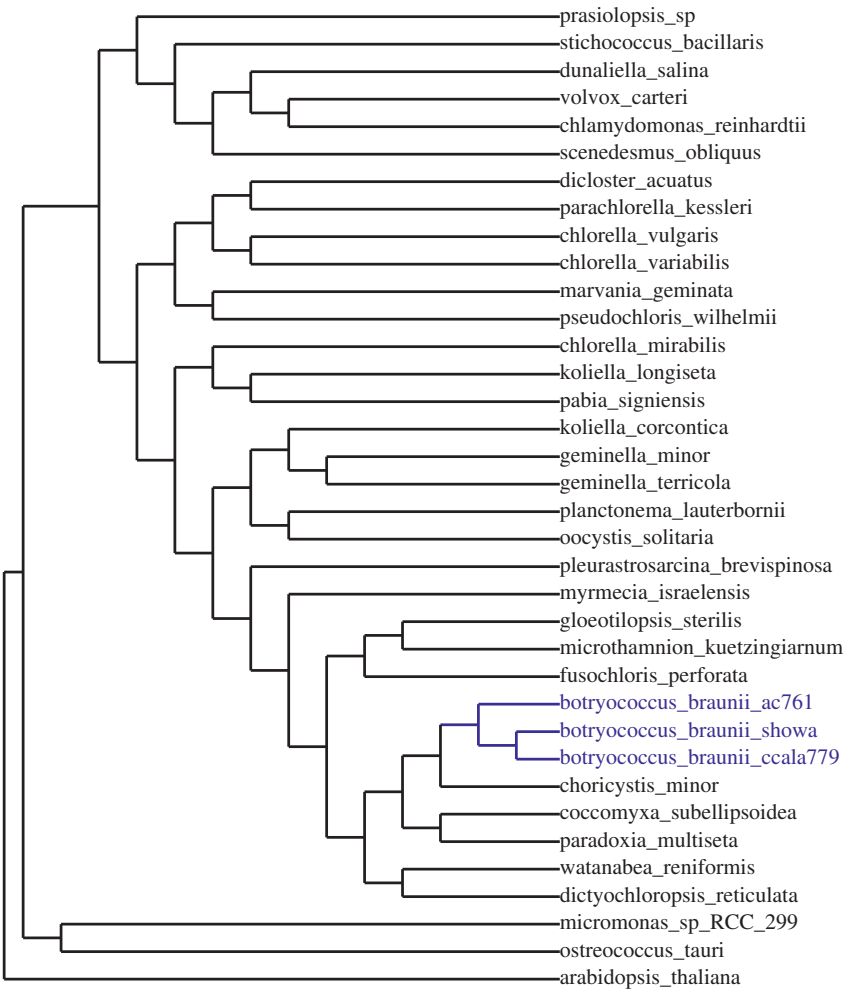


Figure 2: Phylogenetic placement of *CCALA778 and AC761 relative to other microalgal species, based on chloroplast proteins.*

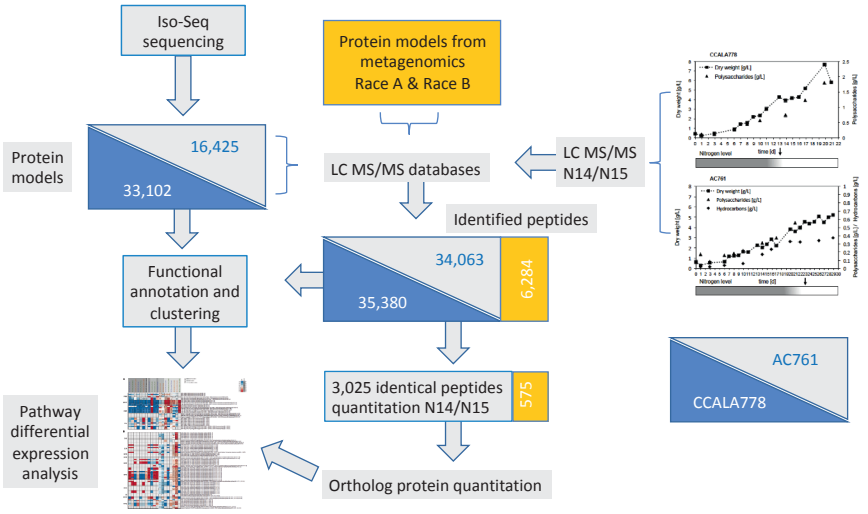


Figure 3: Using Iso-Seq sequencing we identified 33,000 protein models for CCALA778 and 16,000 protein models for AC761. B) The protein models were used to create hypothetical peptides for an LC-MS/MS database. Consecutive proteomics analysis on various conditions validated the protein models. These protein models were assigned a biological function using CrowdGO (chapter 4). C) Using 14N and 15N labelling of the peptides, we distinguished the CCALA778-based and AC761-based peptides. This allowed us to compare the rations between the different races for the various sampled conditions. A thousand proteins were identified to have peptides occurring in both CCALA778 and AC761. These were functionally analyzed using manual curation. Together with the peptide ratios from the quantitative proteomics analysis the thousand proteins were used to analyze several key pathways between the two races to identify key differences and link protein expression to hydrocarbon or exopolysaccharide production.

3.3 Creating the initial protein models

We used Iso-Seq RNA sequencing to identify *Botryococcus braunii* transcripts for both strains, and predicted protein models using MAKER. This resulted in 33,000 protein models for CCALA778 and 16,000 protein models for AC761 (Figure 3). All protein models were assigned Gene Ontology (GO) term annotations using CrowdGO (chapter 4), summarized in table 1. To aid further analysis of the protein models, they were clustered based on 50% or more sequence similarity. These clusters are considered similar enough to present the same protein model with a defined functional entity for further biological analysis.

Table 1: A summary of the Iso-Seq protein functional annotations. A full code enzyme is an enzyme with four codes in the enzyme code nomenclature

	Total
Annotations	391,464
Proteins with a GO term annotation	50,330
Unique GO terms	1,480
Unique enzymes	462
Unique full code enzymes	363
Proteins with full code enzyme	5,553

3.4 Quantitative proteomics analysis

We then performed quantitative proteomics analysis to compare the two *Botryococcus braunii* strains. Races A and B were grown in bioreactors at different conditions and isotopically labeled with either ^{14}N or ^{15}N KNO_3. For comparison, differentially ^{14}N /^{15}N labeled races A and B stemming from the same growth condition were mixed. The rationale for mixing was to quantitatively compare peptides that are identical between race A and race B, in order to measure and relate expression of conserved key metabolic pathways. For proteome analysis the ^{14}N and ^{15}N labelled samples were mixed on equal chlorophyll content and fractionated via SDS-PAGE. From each lane 36 bands were excised and digested tryptically as previously described [225]. In total, 689 samples were analysed by liquid coupled mass spectrometry (LC-

MS). Using the 50,000 IsoSeq-based proteins for the peptide database, a total of 70,081 unique peptides were identified. 35,375 and 34,056 peptides mapped to the race A and race B databases, respectively. 6,284 peptides mapped to the bacterial database derived from metagenomics. For peptide quantification pyQms was used [226]. Retention-time (RT) alignment between all LC-MS/MS runs and enhancement defining RT windows was performed using piqDB earlier [227, 228]. Peptide ratios were calculated for 12,239 peptides. These peptides mapped to 2,407 protein groups in race A, 2,708 protein groups in race B and 575 protein groups in the bacterial database. 1567 peptides were identified and quantified in both races A and B. Due to enhancement via piqDB it was possible to gather quantitative information for some peptides identified in both race A and B, providing ratios and increasing the number of peptides usable for quantitative comparison. In total, 3,025 peptides with a ratio between race A and B were identified. Those mapped on 1,832 protein groups in race A, 1641 protein groups in race B and 322 protein groups in the bacterial database. The quantitative data are displayed as ratios of the absolute number of peptides measured in race A over absolute number of peptides measured in race B, absolute amount in light divided by absolute amount in dark, and absolute amount of nitrate deficient divided by absolute amount in nitrogen presence.

The 3,025 peptides shared between race A and race B belonged to 972 protein clusters as identified earlier based on the IsoSeq-based proteins, and these were used for further comparative analysis. To get a higher resolution of their protein functions we manually annotated these proteins with one defined function, using CrowdGO annotation, BLAST hits, and PFAM hits as a starting point. The manual annotation results are summarized in table 2.

Table 2: A summary of the manual annotations for the orthologs between CCALA778 and AC761. Displayed are the total proteins with an annotation, the amount of unique functions used to describe the proteins, and the top 10 categories of these functions.

	Total
Total proteins	972
Unique functions	841
Protein metabolism	79
Ribosome	68
Translation	55
Lipid and fatty acid	48
Amino acid metabolism	45
Stress response and redox reactions	43
Endomembranes and vesicular trafficking	41
Photosynthesis	41
General metabolism	38
Carbohydrate and sugar metabolism	38

We linked these proteins to eight peptide ratios: light against dark and nitrate against no nitrate for both strains, and early log phase, mid-late log phase, stationary phase I, and stationary phase II between both races. As depicted in figure 5, many enzymes of the respective key metabolic pathways, namely the GDP-L-fucose biosynthesis pathway and methylerythritol 4-phosphate (MEP) pathway could be comparatively quantified. Additionally, we further analysed the core metabolism of *Botryococcus braunii* (Figure 6).

3.5 Fucose synthesis

As described in Figure 1, race A (CCALA778) produces more polysaccharides than race B (AC761). With our proteome data we were able to identify members of the GDP-L-fucose biosynthesis from GDP-L-mannose pathway. In the first

step of this pathway, phosphomannomutase (PMM) converts D-mannose-6P to D-mannose-1P. Then a mannose-1-phosphate guanylyltransferase (GMMP) combines the D-mannose-1P with GTP to produce GDP-D mannose. After that, GDP-D-mannose is converted into GDP-4-keto-6-deoxy-D-mannose by GDP-D-mannose 4,6-dehydratase (GMD), which is then used by GDP-L-fucose synthase to produce GDP-L-fucose.

The peptide ratios show a differential expression of these enzymes between race A and race B (Figure 4). PMM was about 1.5-1.8-fold more abundant in the early logarithmic phase in race A compared to race B but is diminished in abundance as the cell grows. GMMP and GMD were much more abundant in race A when the cells were in the early (4 fold) and mid (4 to 8-fold) logarithmic phase, and the stationary phase (4 fold). Similar abundance changes for GMPP and GMD were found when we compared light and dark grown cells. Moreover, higher protein amounts of GMPP and GMD were detected in race A than in race B when comparing the non-nitrogen. GDP-L-fucose synthase showed varying protein abundances, overall being slightly more abundant in CCALA778 compared to AC761 during the early and mid-logarithmic phases, and 8-fold more abundant in race A when the cells were grown in dark conditions. These data clearly indicate that in the *B. braunii* race A (CCALA778) the enzymes involved in GDP-L-Fucose biosynthesis were significantly more abundant than in race B (AC761), correlating with the increased polysaccharide production in race A (Figure 1).

3.6 MEP pathway

Previous work on *B. braunii* showed that the specific hydrocarbons synthesized by the race B (Showa) are the botryococcenes, which are triterpenoids derived from the MEP pathway [214, 229]. Biosynthesis of linear triterpenoid hydrocarbons occurs via the farnesyl-diphosphate, which occurs as an intermediate in the MEP pathway. First, two molecules of farnesyl-diphosphate are condensed by squalene synthase to form prequalene-diphosphate. Then, in the presence of NADPH, squalene-synthase like 2 (SSL-2) forms squalene, and squalene-synthase like 3 (SSL-3) synthesizes botryococcenes. The extracellular liquid hydrocarbons in race B are usually methylated botryococcenes. Its methylation steps are performed by S-adenosylmethionine (SAM)-dependent methyltransferases (TMT1/2 for squalene, and TMT3 for botryococcenes).

We were able to identify members of above pathway for both race A

(CCALA778) and race B (CCALA778) and observed several key differences in their proteomics ratios (Figure 4). DXR (1-deoxy-D-xylulose-5-phosphate reductoisomerase) and ISPG (4-hydroxy-3-methylbut-2-enyl-diphosphate synthase) were highly expressed in race B compared to race A during early logarithmic growth phase (16-fold). Two other components of the MEP pathway, 2-C-methyld-D-erythritol 2,4-cyclodiphosphate synthase (ISPF) and 4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase (SIPH), were 3-3-fold more abundant in race B compared to race A during the mid-logarithmic and stationary phase.

These data indicated that in race B (AC761) the gene products of the MEP pathway were overall more abundant, likely to provide precursors for botryococcenes.

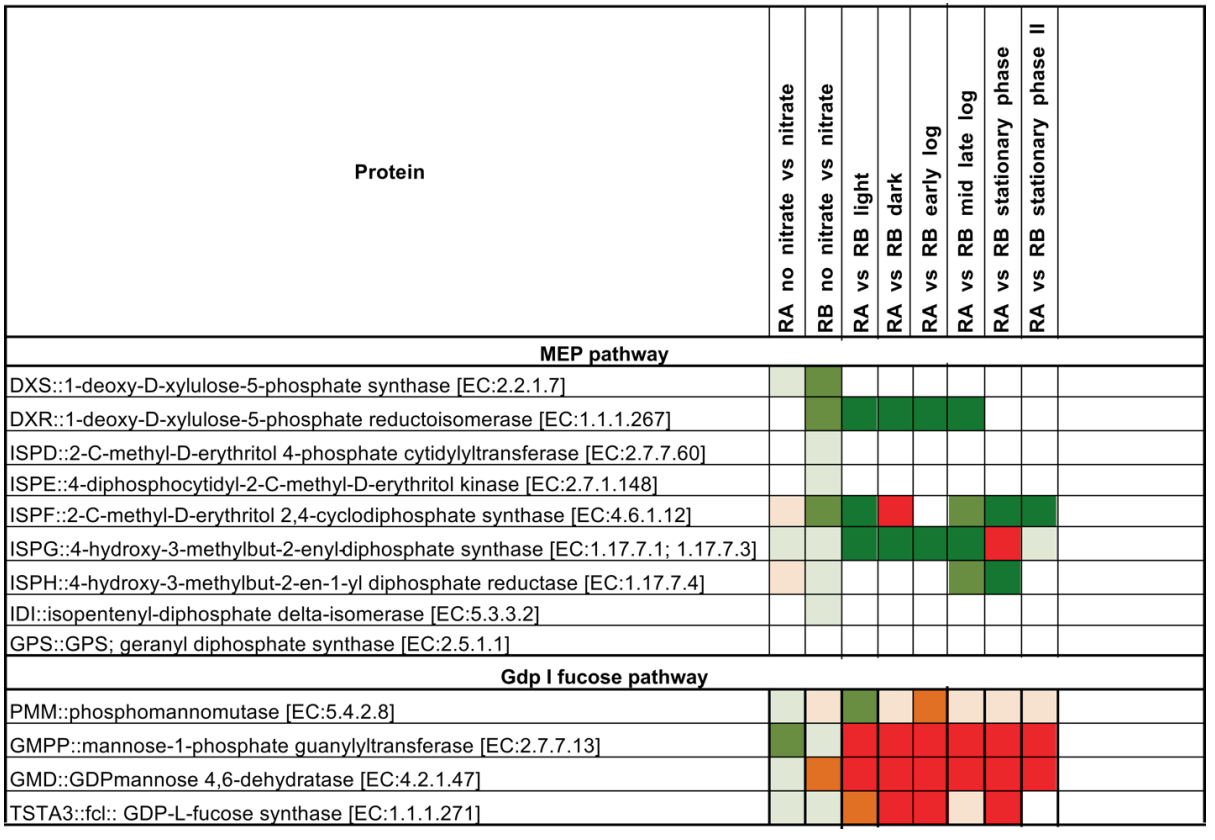


Figure 4: Overview of the protein ratios derived from quantitative proteomics, based on 8 growth conditions. Green is a positive ratio, red is a negative ratio, and white is a neutral ratio

3.7 Proteomics analysis of core metabolism

We identified many proteins in the core metabolism of *B. braunii* for both race A (CCALA778) and race B (AC761). These are summarized in Figure 5.

In the Calvin-Benson cycle, carbon dioxide converts to glucose and other compounds. We were able to detect a number of enzymes involved in the cycle, including RuBisCO, phosphoglycerate kinase, fructose-bisphosphate aldolase, transketolase, ribose-5-phosphate isomerase and phosphoribulokinase. The chloroplast encoded large subunit of Rubisco was more than 2-fold increased from early, mid-log and stationary phase in race B (Figure 6). Transketolase peptide levels were found to be 2-fold more abundant in race A compare to race B in the stationary phase. Fructose-bisphosphate aldolase displayed a more than 2-fold down regulation in race A in the stationary phase but was over 2-fold up regulated in the stationary phase II.

For glycolysis we found proteins corresponding to hexokinase, glucose-6-phosphate isomerase, fructose-bisphosphate aldolase, phosphoglycerate kinase, triose-phosphate isomerase, phosphoglycerate kinase, enolase, pyruvate kinase and phosphopyruvate carboxylase. Triose-phosphate isomerase was upregulated in race A compared to race B in most of conditions tested, with the most prominent increase of over 2-fold seen in the early log phase. Enolase was 3-fold up regulated in race A compared to race B during light conditions, and a little over 3-fold down regulated in the early log phase.

In the citric acid cycle, we identified citrate synthase, aconitate hydratase, isocitrate dehydrogenase, malate dehydrogenase and ATP citrate lyase. The latter was over 2-fold down regulated in race A in the light, dark and the different growth stages.

For the fatty acid biosynthesis pathway we identified acetyl-CoA carboxylase, 3-oxoacyl-ACP synthase, 3-oxoacyl-ACP reductase, enoyl-ACP reductase and acyl-ACP desaturase. Acetyl-CoA carboxylase was 3-fold down in race A compared to race B in dark conditions. 3-oxoacyl-ACP synthase showed a 2-fold down regulation in the light as well as in the early log phase, while 3-oxoacyl-ACP reductase displayed a 3-fold down regulation in the early log phase.

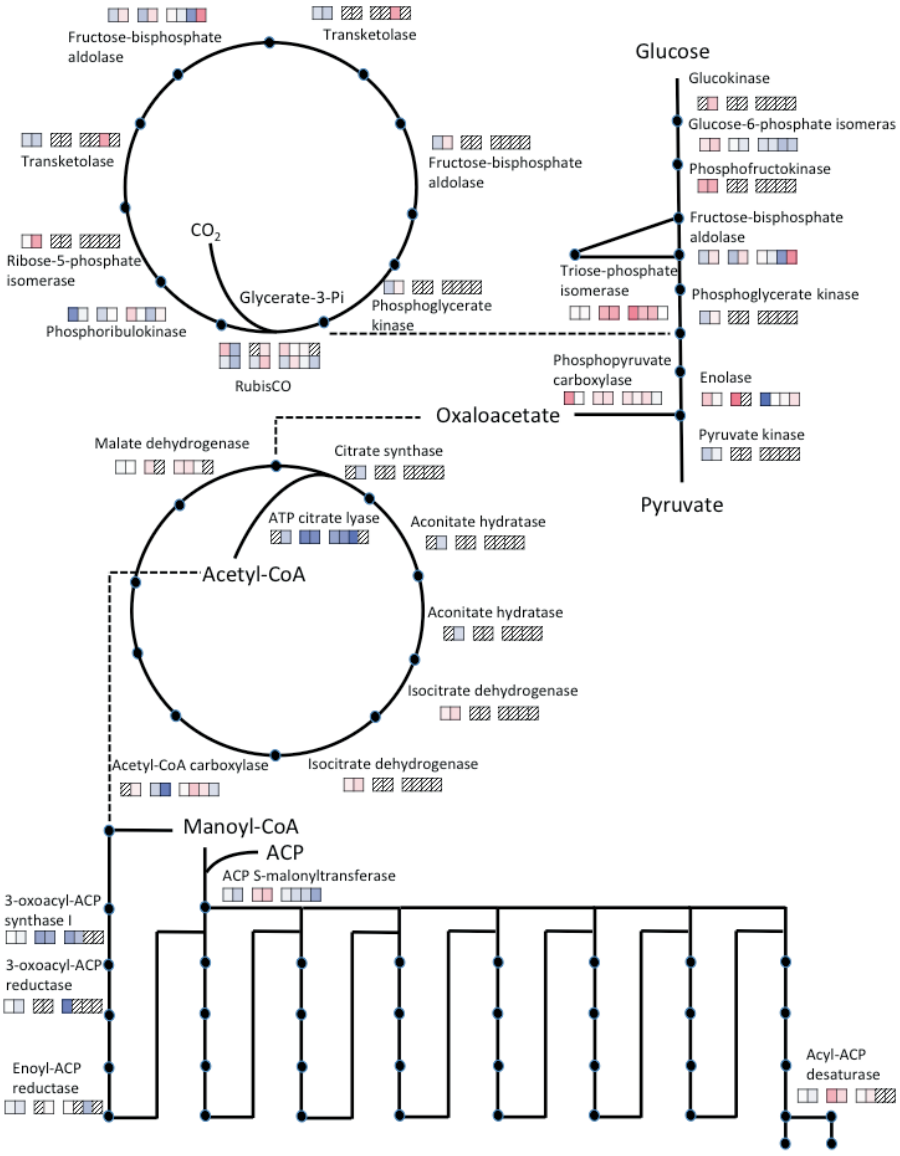


Figure 5: An overview of the protein ratios between the race A (CCALA778) and race B (AC761) strains. The same growth conditions are analyzed as in figure 4. Red indicates more abundance for AC761 orthologs, and blue indicates more abundance for CCALA778 orthologs.

Discussion

This paper aimed to unravel why different *botryococcus* braunii races produce different products in large quantities, namely exopolysaccharides or hydrocarbons. We used a proteomics-based workflow to overcome the challenges of genome assembly, gene prediction, and differential expression (Figure 2). Using the workflow, we were able to compare the two different strains CCALA778 (race A) and AC761 (race B) by looking at commonly shared peptides. Based on peptide ratios between the orthologs of the two races performed a differential expression analysis. This was enhanced by the use of functional information from CrowdGO, and the manual curation of 972 protein clusters which contain orthologs in both CCALA778 and AC761. With this approach we were able to study key differences in the metabolism of the two strains. We found that the race A CCALA778 has several key enzymes overexpressed in the GDP-L-fucose biosynthesis pathway (Figure 5), correlating to its abundant production of exopolysaccharides (Figure 1). In contrast, we found that race B AC761 has several key enzymes overexpressed in the MEP pathway (Figure 5), correlating with its abundant hydrocarbon synthesis (Figure 1). Additionally, we saw differences in protein expression between the two races in the Calvin Benson cycle, glycolysis, citric acid cycle, and fatty acid biosynthesis pathway.

These differences in expression indicate that while both are *B. braunii* strains, and share a high amount of orthologs, there are numerous key differences in their metabolic behaviour. While their product formation suggests enzymes related to hydrocarbon and exopolysaccharide production should be affected, as validated by our proteomics results, there are also vast differences on a core metabolism level. More extensive research on the metabolism of *Botryococcus braunii* should be done to unravel more of its metabolic differences. While in this study we only looked at differential expression in orthologues, follow-up studies could look at differences in the qualitative proteome.

Finally, this paper is a successful demonstration of proteomics and high resolution of protein function information to perform a comparative genomics study of non-axenic microalgal strains, allowing us to go from metagenomic sequences to the identification and comparison of microalgal orthologs and their functions. In the future, similar approaches should be used for likewise large-scale studies of non-axenic microalgae.

Chapter 7

General discussion

1.1 - The SPLASH project

The study of this thesis was done in light of the Sustainable Polymers for Algae (SPLASH) project [230]. This EU consortium was funded to study the oleaginous microalgae *Botryococcus braunii* because of its potential for producing large amounts of hydrocarbons and polysaccharides, which are excreted in large quantities out of the cell [204, 216]. A large part of the SPLASH project was the genomics study, where we aimed to characterize the genome of *Botryococcus braunii* using omics data, specifically for the identification of key proteins involved in hydrocarbon and polysaccharide production. For this, the aim was to apply comparative genomics between the hydrocarbon producing *B. braunii* race AC761, and the polysaccharide producing race CCALA778. With the comparative genomics study, we were able to find key differentially expressed proteins in the non-mevalonate pathway and the L-GDP-Fucose biosynthesis pathway, related to hydrocarbon and polysaccharide production respectively. These findings provided us testable hypotheses for the engineering of these proteins in other, faster growing microalgae such as *Chlamydomonas reinhardtii*.

Genomics, however, was only one part of the SPLASH project. The multi-disciplinary consortium was set up to investigate many aspects of *Botryococcus braunii*, with the main studies being:

1. Development of innovative cultivation and downstream concepts for improved growth, product enhancement, and integrated recovery of polysaccharides and hydrocarbons.
2. Product development and testing
3. Process demonstration at pilot scale
4. Process integration, sustainability assessment, and market analysis

Each of these studies is important in their own right, and a few can utilize any results from genomics studies.

1.2 Genomics for the development of cultivation and downstream concepts for improved growth and product enhancement

An important part within the SPLASH project, and a widely studied factor in microalgal research, was trying to find innovative ways of cultivating and growing *B. braunii* and extract its hydrocarbons and polysaccharides. For the first goal, growth media, day and night cycles, and growth temperatures,

amongst others were optimized. This resulted in 3-fold increased polysaccharide and 2-fold increased biomass production. These numbers can be increased in two ways: first, by analysing genetic effects on specific growth media and environmental conditions and attempting to optimize these further based on the genetic responses. And second, as discussed in the introduction, there are metabolic constraints that limit the potential growth, hydrocarbon, and polysaccharide production of microalgae. Finding these constraints and possibly circumventing these by the use of metabolic engineering or synthetic biology will provide a boost to the growth and production capabilities of *B. braunii*.

1.3 Genomics to aid in process demonstration at pilot scale

An important question for microalgae, is how they can be cultivated in large open pond algae farms for commercial purposes. As part of SPLASH, several 'hotspots' were found that increase the cost for commercializing microalgae, including the need to increase the amounts of hydrocarbons and polysaccharides being made in a pilot scale setup. This subtask did not succeed in developing a pilot scale setup able to produce large amounts of hydrocarbons or polysaccharides. While no definitive reasons are given, one of several reasons might be that *B. braunii* lives in communities with a high number of bacteria. It is hypothesized these consist of symbiotic, parasitic, and mutualistic relationships, and that some bacteria enable *B. braunii* to produce large quantities of hydrocarbons and polysaccharides. For these questions, genomics can be used to identify key proteins that might interact with proteins or compounds coming from or going to bacterial sources. One such technique that can be used is that of mathematical modeling, based on the metabolic capabilities of the microalgae and those of the bacteria.

1.4 Genomics to aid process integration, sustainability assessment, and market analysis

Part of any big project like SPLASH is a market analysis, to assess if the commercialization of a product is viable, and if not, how it is to be made viable. It was concluded that, as was predicted, hydrocarbon and polysaccharide production with *B. braunii* is the most viable. While currently not ready to put on the market, the main conclusion of the final report was:

"It can be concluded that the in SPLASH developed cultivation and milking technologies have the potential to operate economically, in particular for EHC,

provided that all the involved processes will be optimized and the targeted products are of higher than commodities value." [230]

For this purpose, genomics is able to aid in product synthesis optimization. In particular, metabolic engineering of *B. braunii* can lead to a higher production of specific hydrocarbons and polysaccharides, or increased growth of the algae. This would lead to a more favourable market analysis for *B. braunii* by decreasing the cost for cultivation relative to hydrocarbons and polysaccharides produced.

All three topics can potentially use predictions based on genomics studies to create testable hypotheses. These can be gained through the identification of key enzymes, the study of pathways, or by utilizing mathematical models. What they have in common, is that at their core they require a good functional annotation of their proteins. As discussed in the introduction, improvements need to be made in the prediction of microalgal protein functions. The study of microalgal protein functions in this thesis is therefore not only for the benefit of the genomics study in SPLASH, but also in a broader sense for other algae projects.

2 - Goal of the thesis

With the SPLASH project in mind, the goal of this thesis was to develop and use bioinformatics tools and pipelines to increase our understanding of oleaginous microalgal cell factories. During my thesis I mainly worked on the improvement of protein function annotations. Genome annotation of the basidiomycota yeast *Cutaneotrichosporon curvatus* was used as a proof of concept for my developed function prediction tool and manual curation pipeline (**chapter 5**), and in **chapter 6** I apply these for the comparative genomics of two races of the *Botryococcus braunii* microalgae. Here I will mainly focus on protein function annotation, how these annotations are connected to microalgal cell factories, and how they can be further improved. Finally, I discuss how comparative genomics can be used for protein farming in microalgae.

Knowing the functions of all proteins of an organism, its functionome, is a core requirement to understand the capabilities of the metabolism of an organism at a genetic level. There are several ways of retrieving protein functions: through wet-lab experiments, manual curation of protein features, or by

computationally predicting them. Figure 1 visualizes a workflow on how an ideal case scenario of a microalgal cell factory functional genomics study, such as that for the SPLASH genomics, is connected to these various stages of annotation and what each step provides towards more knowledge of cell factories.

Step 1: High-throughput annotations

Normally the first step in functional genomics for the study of microalgal cell factories is high-throughput predictions. This forms the basis for further research, and a high accuracy and coverage of annotations allows for easier follow-up studies and a better understanding of the metabolism of an organism. As explained in the introduction, there is a wide array of methods that are able to predict in high-throughput which Gene Ontology (GO) terms belong to a protein [64], most of them falling in the sequence similarity and machine learning categories. Many of these methods participated in the Critical Assessment of protein Function Annotation (CAFA) competition [56, 182], where they aim to correctly predict the GO terms belonging to proteins with an unknown function. The most recent finished CAFA competition compared 126 prediction methods from 56 research groups, which are assessed on 3,681 proteins from 18 species. Because of its scale and competitiveness, any functional genomics research should be utilizing one or more of the well-performing methods participating in CAFA.

The Critical Assessment of protein Function Annotation

Because predicting protein functions is an important subfield in bioinformatics, there is a wide array of ever improving prediction methods available. These are hard to compare, because they all use different data sets and different evaluation metrics to benchmark their performance on. The CAFA competition aims to find the top performing protein function prediction methods by benchmarking them using the same evaluation metrics. The first CAFA results got published in 2013 and evaluated the performance of 54 methods based on 866 proteins from 11 species [56]. Predictions were initially done on 48,298 proteins, but only the proteins that were experimentally validated over the course of 15 months after the submission deadline were used for the evaluation of the methods. For this edition of CAFA, only biological process and molecular function GO terms were considered. BLAST [60] and Naïve predictions were used as a baseline, where BLAST predictions used the top sequence similarity hit to transfer GO terms to the target protein and its

sequence similarity as a confidence score, and Naïve predictions assigned a confidence score to each GO term based on their relative abundance in the SwissProt database. The results of the first CAFA showed Jones-UCL and Argot2 [2] as the winner and runner up for both biological process and molecular function predictions. As a result, I used FFPred2 [167] from UCL-Jones and Argot2 for the duration of this thesis.

However, apart from showing the top performers, there were multiple notable results to be observed in the study. BLAST was outperformed by 33 methods in the molecular function category, and 26 methods in the biological process category. Also, for all methods including BLAST, biological processes were harder to predict than molecular functions. The authors hypothesize that this is likely due to biological processes being more abstract in their function compared to molecular functions, making them harder to predict by looking at straightforward amino acid conservation [56].

Furthermore, the authors looked at performance differences between 'easy' and 'hard' to predict proteins. Proteins with 60% or more sequence similarity to an experimentally validated protein were classified as easy, and others as hard. Unsurprisingly, BLAST had a tougher time predicting hard proteins than easy proteins. However, the top performers of CAFA showed no significant difference in predicting easy or hard proteins. This hints that the state-of-the-art prediction methods are good at utilizing multiple sources of data and were able to compensate for a low sequence similarity to proteins with a known function.

The second edition of CAFA was published in 2016. 126 methods participated and were benchmarked on 3,681 proteins from 18 species. Most of the general results were similar to those of the first edition, however the top performers of the second edition outperformed the top performers of the first edition. This indicates an improvement over time for protein function prediction methods. However, it has to be taken into consideration that the increase of available data in databases such as SwissProt [55] leads to better training of prediction methods, and more information to be used for predicting protein sequences.

The first edition of CAFA contained detailed information on the techniques used for predictions by each team. Looking at the top five competitors in this edition, we get a varied list of techniques used:

1. Jones-UCL: profile-profile alignments, sequence properties, protein interactions, gene expression, literature, machine learning, orthology
2. Argot2: sequence alignments, sequence-profile alignments
3. Pannzer: Sequence alignments, profile alignments, orthology, paralogy [231]
4. ESG: sequence alignments [232]
5. PDCN: profile-profile alignments, sequence-profile alignments [233]

In conclusion, BLAST-based protein annotation, which is still considered an acceptable method of GO term annotation by many, is shown to be vastly outperformed by state-of-the-art prediction methods and should not be used directly for these kinds of purposes. Because of its scale and competitiveness any functional genomics research should be utilizing one or more of the top-performing methods in CAFA.

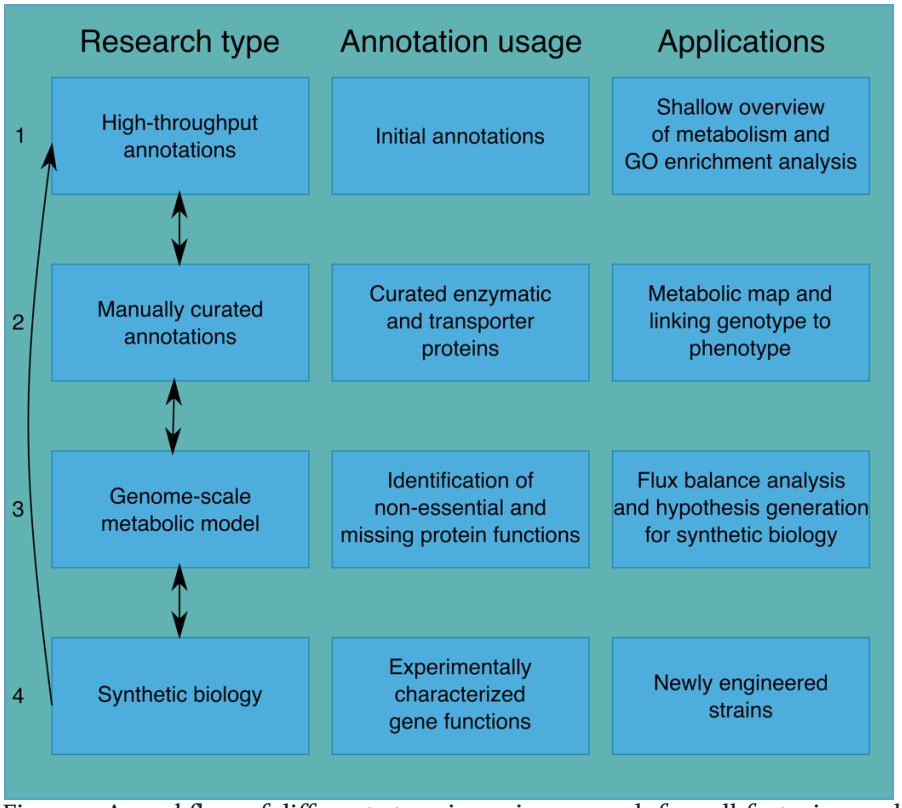


Figure 1: A workflow of different steps in omics research for cell factories, and how they are interconnected by protein function annotations. 1) High-throughput annotations form the core for any further research, but is limited in its usage. 2) Manual curations improve upon these high-throughput annotations, providing the first detailed metabolic map. These curated proteins can be used as a template for future high-throughput annotations of other species. 3) The manual curations are used to create a genome-scale metabolic model of the organism, describing the metabolic reactions using mathematical equations. These can be used to form in silico driven hypotheses for metabolic engineering experiments. Removing, introducing, or over expressing genes in the model gives more information on the protein functions, which can be used to update the manually curated proteins. 4) Metabolic engineering experiments are done to test the hypotheses. These lead to new strains of the organism. Characterization of the changes introduced by removing, introducing, or over expressing genes validates its function. These validated functions can be used to improve the genome-scale metabolic model, and as a template for future high-throughput predictions.

Merging high-throughput annotations with CrowdGO

As discussed in the introduction, in an ideal case scenario the GO terms of several prediction tools are combined to form a more accurate and complete representation of protein functions. The CAFA results show that from the top five performers, none of them use the same combination of techniques. This implies that there are vast improvements to be gained from combining the predictions of different methodologies, as they might be complementary to each other. For this purpose, **chapter 4** describes CrowdGO: a tool able to merge the predicted GO terms from different prediction tools, resulting in significantly improved GO annotations. In **chapter 5 and 6** we used two of the top performing methods of the most recent CAFA: FFPred2 from the Jones-UCL team, and Argot2. To supplement them, we used the widely used prediction tool InterProScan, which works with sequence-profile alignments, shown to be a top performing technique in CAFA. The predictions of these methods were merged with CrowdGO. This has led to significantly more accurate predictions (**chapter 4, figure 3**), and we further used this way of protein function annotation in both **chapter 5 and 6** for functional genomics.

Shortcomings of high-throughput annotations

However, high-throughput prediction methods are limited. These GO term predictions are often very general, or in low amounts. Their predicted GO terms are either shallow or have a low coverage. For example, GO:0016298 represents any lipase activity, and has ten directly related child terms, such as triglyceride lipase activity and phospholipase activity. In most cases, the GO term for lipase activity will be predicted instead of one of the more downstream terms. As a result, most high-throughput annotations can only be used as a general guideline for the function of a protein. Because CrowdGO merges GO terms from other prediction methods, it has the same limitation.

A straightforward solution for getting more specific GO terms is by improving the input predictions for CrowdGO. New CAFA2 predictors outperformed the top predictors of CAFA1, and during the end of my thesis improved versions of Argot(2.5) and FFPred(3) became publicly available. The first improvement to be made is to use the top-performers of the second CAFA edition and see if with this input CrowdGO is able to predict more specific GO terms. Also, the enzyme prediction tool EnzDP became available. This is a prediction method specifically for predicting enzymes, resulting in highly specific enzyme annotations. Applying EnzDP on the oleaginous yeast *Cutaneotrichosporon*

curvatus from **chapter 5** resulted in an overlapping prediction of 75% with the enzymes involved in the manually curated triacylglycerol synthesis pathway (Supplementary Table 1), which is a high recall for high-through methods. Because we are mainly interested in enzymes for our metabolic research, EnzDP is an interesting option.

Another thing to take into account is that most high-throughput methods use SwissProt proteins as a reference for their predictions. For example, CrowdGO uses a training and test set derived from SwissProt proteins to train its parameters on. Because SwissProt is exclusively manually curated it contains a relatively low amount of proteins compared to for example the high-throughput annotated TrEMBL database. This means that a training and test set for CrowdGO contains a limited amount of proteins, consisting of predominantly a select few species: humans and other mammals, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Oriza sativa*, and various bacteria. These lead to a biased set of proteins used for the testing of CrowdGO and other protein function prediction methods. As such, while CrowdGO performs well given the circumstances of chapter 4, it is not a given that it will perform as well on the proteins of non-model species such as microalgae, which are not represented in the top 50 SwissProt species. Additionally, because CrowdGO works largely based on pattern recognition of training data derived from SwissProt, there could be a chance that it has a hard time correlating patterns of non-model species to highly specific GO term predictions.

Hopefully in the future predicting highly specific GO terms will become more consistent. However, high-throughput predictions currently are only applicable for a general overview of an organism's metabolism and function enrichment analysis, if expression data is available. For any further analysis, the high-throughput annotations will most likely be too shallow, inaccurate, and incomplete.

Step 2: Biocuration of protein annotations

The next step in the research for microalgal cell factories is functional genomics through manual curation of proteins involved in a microalgae's metabolism. Using high-throughput predictions we are able to get most of the conserved metabolism of microalgae, due to sequence similarity with experimentally characterized proteins of *Arabidopsis thaliana.* However, most microalgal metabolic proteins do not have well-characterized homologous

proteins. This subset of proteins should be manually curated. For this step, the high-throughput annotations are an excellent starting point, especially if they are relatively accurate to begin with.

Step 2.2: Using CrowdGO in a biocuration pipeline

The most well-known database of manually curated proteins is SwissProt, part of the UniProt database. As discussed in the introduction, SwissProt uses an extensive pipeline for the manual curation of its proteins (**chapter 1, figure 2**). This database is regarded as the gold standard for protein functions and is used by most prediction methods including CrowdGO. However, its protein curation is a slow and laborious process, as is reflected by comparing the number of proteins to those of the non-curated UniProt TrEMBL database: 550 thousand proteins in SwissProt against 115 million in TrEMBL. Because of this reason, we implemented CrowdGO in a manual curation pipeline, in addition to BLAST, PFAM, and HMMER searchers for the curation of 1,700 proteins in **chapter 5 and 6**. The annotation process of the 700 proteins in chapter 5 took the equivalent of two weeks for one person. Additionally, the manually curated proteins were used to fully characterize triacylglycerol (TAG) synthesis of *Cutaneotrichosporon curvatus*, with only one enzyme initially missing. This reflects a high and accurate coverage of manually curation using this pipeline.

In **chapter 6** we biocurated a 1000 proteins of *Botryococcus braunii* in a similar manner. However, microalgae have a complex genomic architecture, and complicated sequencing efforts are required to assemble their genomes with high quality. Part of the genomics study of SPLASH was to create a high-quality genome and gene set of *B. braunii*. This wasn't possible before a large amount of long read sequencing was done, using PacBio and HiSeq sequencing. Furthermore, gene curation was done using the widely used MAKER tool but required extensive biocuration by overlaying RNA IsoSeq sequences on the genome, and visually inspecting and correcting these genes in CLC workbench. The process of attaining a high-quality genome for *Botryococcus braunii* CCALA78 and AC761 and their respective gene sets took place over the entire duration of the SPLASH project. While we managed to achieve a high-quality genome for *B. braunii*, the majority of other microalgae have lower quality genomes. Additionally, the phylogenetic tree we constructed in **chapter 6 figure 2** shows *B. braunii* species are not closely related to extensively studied microalgal model species such as *Chlamydomonas reinhardtii, Volvox carteri, or Ostreococcus taurii*. This means we did not have a well-curated genome to use as a reference for the manual curation of proteins. Due to the lack of

comparative genomics this manual curation process was slower than that of *C. curvatus*, resulting in two people curating the protein functions for a month to achieve the same accuracy as for those of *C. curvatus*.

Incorporating CrowdGO in biocuration

For the foreseeable future the SwissProt database will likely be the gold standard [55], and it is unlikely they will adopt our manual curation pipeline. However, UniProt has eight on-going biocuration projects for specific organisms [234]. Cooperation with the UniProt consortium to start a microalgal biocuration project would be a tremendous boost to the study of microalgae, and efforts should be made from the microalgal community to start such a project. In this case, the manual pipeline used for the curation of *Botryococcus braunii* metabolic proteins can be used as a starting point for such a project. *Botryococcus braunii* might not be the ideal representative for microalgae as a UniProt biocuration project, however their proteins can be incorporated for other oleaginous microalgae. For example, the manually curated *B. braunii* proteins for the non-menavolate and GDP-L-Fucose synthesis can serve as a conserved reference to other oleaginous microalgae. A more feasible species to adopt for a microalgal biocuration project should be easier to manipulate in the wet-lab for the purpose of any required validation experiments. One species that comes to mind is *Chlamydomonas reinhardtii* [156], and another one is the oleaginous *Nannochloropsis gaditana* [235] which due to its properties might reflect the interest in microalgae better.

One way SwissProt can benefit from the work in this thesis is by the annotation of TrEMBL proteins [55] using CrowdGO. The SwissProt curation pipeline uses TrEMBL proteins as a starting point, and therefore more and better information on their function will improve the accuracy and speed of curation. TrEMBL proteins with existing GO term annotations can be updated by merging them with CrowdGO, or the entire database can be revamped de novo using a standard set of prediction tools such as InterProScan [61] and selected CAFA methods, and merged using CrowdGO as described in this thesis. Given the significant increase in accuracy, there is no downside to applying the tool.

In conclusion, CrowdGO can play a positive role in the biocuration of yeast and microalgae, as shown in **chapter 5 and 6**. Biocuration of these proteins has led to an increased resolution of metabolic information on these species. Finally, if third party manual annotations are fed back into protein databases such as SwissProt, they can serve as reliable blueprints for the high-throughput

annotation of other proteins. Currently, all high-throughput annotations are based off of SwissProt proteins so any extensions to this database will increase the usability of high-throughput prediction methods.

Step 3: Constructing a genome-scale, constraint-based metabolic model

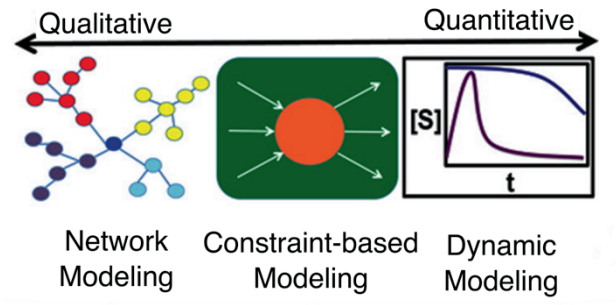


Figure 2: Network analysis, flux balance analysis, and kinetic modeling in a schematic overview from qualitative modeling to quantitative modeling. Adapted from [1].

There are several types of models used to describe microalgal metabolism. Lenka et al describes the major types used so far in the metabolic modeling of microalgae [1]. These can be described using qualitative modeling: a representation of metabolism mostly by visualization and concepts using little to no mathematics, and quantitative modeling: a representation of metabolism using one or more equations. Figure 2 shows how these different types of models fall on the qualitative to quantitative scale.

Network modeling can be done using topological network or graph analysis. These forms of modeling describe the interaction between proteins in the network, and their place in the network. While these models are relatively easy to interpret, they provide little prediction power as to metabolic flows of the organism. For splash we use topological network modeling to describe hydrocarbon and polysaccharide production of *Botryococcus braunii* (**chapter 6**).

Constraint-based models are representations of a biological system in a stoichiometric matrix, with each row representing a metabolite and each column a reaction (**chapter 2, box 2**). Such a model, as mentioned in the introduction, consists of a set of algebraic equations describing the

stoichiometry of the metabolic reactions of an organism, up to genome-scale. A detailed metabolic map of an organism provides a blueprint for creating a genome-scale metabolic model (GSMM) [109, 124, 236]. This is useful for predicting metabolic fluxes under given specific operating conditions and can be used to propose gene knock-outs, knock-ins, and over-expressions.

One technique to analyse constraint-based models is flux balance analysis. Flux balance employs two constraints: the metabolism is assumed to be steady state, meaning no production or degradation of intermediate compounds is possible, and each reaction's flux is given an upper and lower bound depending on thermodynamics and substrate availability. One or more reactions are then set as the objective function, such as triacylglycerol production. With the matrix, the constraints, and the objective function, we now have a linear programming problem, for which the optimal solution is found. Flux balance analysis is the most widely used modeling approach in microalgae, mostly in the form of genome-scale metabolic models. Several of such models exist, mostly for *Chlamydomonas reinhardtii* as described in **chapter 2**.

Finally, dynamic models describe the metabolism of an organism using rate equations and are generally represented in the form of an Ordinary Differential Equation (ODE), or a series of ODE's in the case of multiple reactions. These dynamic models are able to describe the metabolism in much more detail than with network or constraint-based models, however they are highly dependent on accurate knowledge of the reaction order and rate constants. Due to their mathematical complexity and necessary input data, dynamic models are generally used to describe small biological networks in high detail. One example is Tevatia *et al* who describes the relation between growth, nitrogen levels, and lipid accumulation in *Chlamydomonas reinhardtii* using a dynamic model [237].

Because of the widespread availability of stoichiometric data, the direct relationship between protein functions and their metabolic reactions, and the questions that flux balance analysis can answer, genome-scale, constraint-based metabolic models are effective to broadly describe the metabolism of a microalgae.

Step 3.2: Improving protein annotations by using Genome-Scale Metabolic Models

In the genomics research of **chapter 5 and 6** we don't create a GSMM to further study the species. However, it would be the logical next step for both

studies. There are a few automatic GSMM generators, which provide models are relatively correct in a general sense but require manual curation to provide more detail. Even so, these automated GSMM's act as a good starting point for further research. Importantly, these automatic GSMM generators can aid in the annotation of an organism's functionome by identifying which enzymes and transporters are missing. If flux balance analysis can't solve the model it means there are enzymatic or transporter reactions missing, indicating there is a gap in the metabolic map. Additionally, if the *in-silico* phenotype does not match that of the wet-lab, it might also be due to a missing protein or pathway. But in this case, it could also be due to other factors such as wrong equations in the model. In the case of a gap in the model all proteins should be checked for hints of any protein able to fill this gap.

Step 4: Creating new strains

A thorough high-quality genome analysis can lead to actionable knowledge to improve product formation. Given the hypotheses generated by the use of the GSMM, genome engineering techniques such as CRISPR-CAS9 can be used to replace, over express, or introduce new genes in an organism to create strains with better industrophilic properties. These strains are hopefully commercially viable for the production of interesting compounds. Even if not, characterizing the new strains and the effect of the engineered genes provides valuable information on protein function and any metabolic changes. Iteratively, these engineering experiments can be used to further inform the GSMM, resulting in an iterative Design-Build-Test-Learn cycle as discussed in the introduction. Furthermore, any highly reliable information gained on protein function should be documented in a FAIR manner (**section 4**). By documenting the information on protein function, metabolic engineering experiments improve the performance and usability of high-throughput predictions (step 1) for other species by acting as a reference, completing the circle of protein function annotations.

4 - Microalgae as protein farms

4.1 The potential of microalgal protein farms

Microalgae are not the only oleaginous eukaryotes with a lot of potential, and we are not sure if them being a host organism for commercial purposes is optimal due to their metabolic constraints [42]. To make microalgae viable as cell factories, many metabolic constraints need to be circumvented by

metabolic engineering, including potentially its photosynthesis mechanics. Using other organisms as cell factories are possibly more straight forward, however, these often don't have the interesting characteristics of microalgae. One example is the possibility of *Botryococcus braunii* to synthesize large amounts of triacylglycerols and to excrete them, which would mean an easier extraction process of these lipids compared to organisms that don't, potentially reducing the cost of triacylglycerol synthesis. However, *Botryococcus braunii* has slow growth and is difficult to maintain in a steady-state condition due to it living in a community with a vast amount of bacteria [238]. In this case it would make sense to genetically engineer the genes responsible for *B. braunii* triacylglycerol synthesis and excretion to an organism that grows faster and is easier to main, for example the microalgal model species *Chlamydomonas reinhardthii* or a yeast. With an estimated amount of 500,000 to a million microalgae in the wild [239], they would provide excellent sources as gene farms.

4.2 A workflow for mining microalgal proteins

This means that large-scale comparative genomics would be an excellent way to study microalgae, but also other oleaginous organisms such as fungi or cyanobacteria as they could potentially be hosts for microalgal genes. This would allow for the identification of key proteins in both interesting metabolic and non-metabolic proteins, and the cross-linking between microalgae and other organisms. Figure 3 summarizes a mock up pipeline for this approach:

1) Select an oleaginous organism of interest.
2) Retrieve the proteome of the organism.
3) Use CrowdGO to high-throughput annotate the proteome.
4) Manually annotated the proteome of the species.
5) Perform differential expression analysis on as many growth conditions as possible.
6) Gather phenotypic data.
7) Use point 3, 4, and 5 to summarize everything in biological pathways.
8) Store everything in a database for oleaginous organisms. This database is the starting point for hypothesis creation for new strain development using synthetic biology.
9) Retrieve and compare the phenotypic traits of the oleaginous organisms.
10) For the phenotypic traits of interest, select the related proteins and their differential expression analysis.
11) With this information, there is now a selected subset of proteins and their meta data related to a phenotypic trait.
12) Use this subset of proteins to create new strains using CRISPR-CAS9 [41, 240]. Feed its phenotypic data back into the database, and in the optimal case, use the newly developed strain to commercialize a product such as hydrocarbons or polysaccharides.

A comparative genomics pipeline like this is ambitious and poses a few challenges. Different data is handled by different people at different times. This means that there are a multitude of different assembly, gene prediction, protein function prediction, and other analysis strategies used. Furthermore, existing and future data for different species and projects use different environmental conditions and differ in quality. Finally, analysing vast amounts of big data requires computational methods that are able to handle and store this data in an efficient manner.

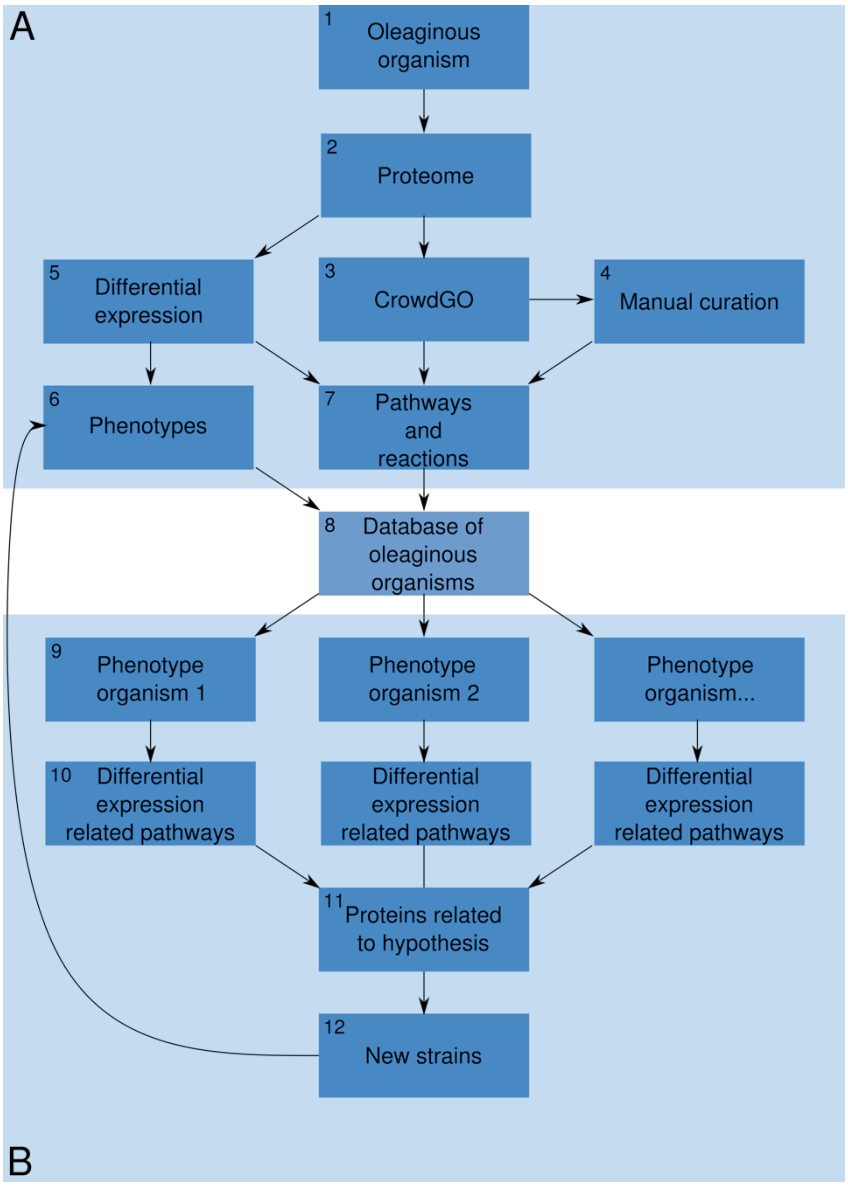


Figure 3: An overview of a mock-up comparative genomics pipeline used to harness comparative genomics for new oleaginous strain development. A) Input data required for the oleaginous organism database, B) utilizing the oleaginous organism database.

4.3.1 Standardization for comparative genomics

One challenge of large-scale comparative genomics research, as described in Figure 3, is the different ways data sets are generated, and the different ways these data sets are analysed. Differences in data generation and analysis produce small to large differences in the results, and most importantly, introduce different errors. These differences make it hard to distinguish between actual biological differences and those introduced by different tools and methodologies. In order to facilitate comparisons between data sets it is necessary to standardize the procedures for data acquisition and analysis as much as possible and be transparent about these procedures. For transparency, the FAIR (Findable Accessible Interoperable Reusable) data principles should be followed [241]. The FAIR principles are guidelines on how to store all generated biological data. This means the data should be stored together with its metadata such as where the samples are from, how the sequences were sequenced, and what data analysis tools were used. Following these principles, analysis and microalgal projects in the future will be a lot easier to perform, and unnecessary research or data generation can be avoided. An example of a pipeline using standardization and FAIR principals is SAPP, a Semantic Annotation Platform with Provenance. This pipeline provides automated structural and functional annotations given a genome, and stores these in a FAIR manner. It is built on the GBOL ontology, a standardized RDF data structure built specifically to make big data interoperable. This has made it possible to, for example, perform comparative genomics on 432 *Pseudomonas* strains [242], and to use protein domains as a fast alternative for sequence-based similarity approaches on functional genomics [243].

The case of the workflow described in Figure 3 requires standardization in the following aspects: genome assembly and error correction, gene prediction, protein function prediction, differential expression analysis, and databases from which to retrieve metabolic information.

4.3.2 Assembly

Assembling a genome is the first aspect of any genomics research, and the hardest to standardize. Due to the differences in next generation sequencing technologies used [244], for example Illumina , PacBio, or 10x sequencing, or due to differences in cultivation and growth of the to-be-sequenced organism, there is no one-rules-all assembly tool out there. There are short read assemblers, long read assemblers, and hybrid assembler that combines long

reads with short reads. Additionally, there are de novo assemblers and reference-based assemblers. On top of that, some assemblers might work better depending on the organism or the quality of the sequencing data. Therefore, it is almost impossible to standardize genome assembly.

What can be done, however, is standardize the quality of a genome or transcriptome assembly to be used in a comparative genomics analysis. BUSCO (Benchmarking Universal Single-Copy Orthologs) is a tool developed to assess the quality of a genome or transcriptome assembly based on its ability to find single-copy orthologous genes that are present in all of a certain taxonomic class [245]. It de novo predicts genes based on the given genome or transcriptome assembly, and returns a percentage of completeness, which is the amount of single-copy orthologs genes it was able to predict given the assembly. In the case for green algae, the orthologous genes are based on the embryophyta, a clade of green plants. In the nearby future it would be useful to create an orthologous group of BUSCO proteins for green microalgae. All genome and transcriptome assemblies to be included in a large-scale comparative genomics study should pass a certain threshold of BUSCO completeness, for example 90% or 95%. A lower amount of completeness means an inability to predict a large amount of proteins correctly.

4.3.3 Gene prediction

The next step is gene prediction. When comparing the metabolism of species on a genetic level we are comparing their proteins and their functions, and how these interact with other proteins and compounds. Therefore, gene prediction is essential. However, even the best gene prediction tools make many errors that can alter the length, reading frame, and hypothetical function drastically. Luckily, unlike with genome assemblies, gene prediction can mostly be done using the same tool. There are only two exceptions: prokaryotic genes are predicted differently than eukaryotic genes, and fungal genes use different parameters for gene prediction than other eukaryotes. For the case of this pipeline, where we are primarily interested in the gene prediction for microalgae and some fungi, it is possible to predict all genes using either BRAKER [246] or MAKER [247], which are shown to be state-of-the-art gene prediction methods. It is even possible to take the combined prediction of genes to get a more complete gene set, but which might introduce a few more false positive predictions.

4.3.4 Function prediction

After gene prediction and translation to proteins, the next step is protein function prediction. It is important that the best function prediction methods are used and that these are standardized cross-species, as these predictions are highly differential between methods as shown in **chapter 4 and 5**.

4.3.5 Differential expression analysis

Apart from using the same differential expression tool and settings for every to-be-compared gene set, it is also important to use the same pre-processing tool to generate the input data. In the case of differential expression analysis this concerns aligning reads to the genes. This can be done with TopHat2 [248], Bowtie2 [249], or STAR [250]. Two standardly used differential expression analysis packages are EdgeR [200] and DEseq2 [251].

4.3.6 Databases from which to retrieve metabolic reactions

The final step of the workflow described in Figure 3 is overlaying the differential expression on pathways. These pathways and other metabolic information can be retrieved from reaction and pathway databases, such as KEGG [69], WikiPathways [71], MetaCyc [72], and more. However, due to the differences in these databases, which are often contradictory, and missing data in some of them, it might be better to include a few databases in order to include as much data as possible. Because the step of analysing differential expression data overlaid on pathways is mostly done without the aid of any further tools, the user can make a well-informed decision on which reactions and pathways to include from which databases.

Concluding remarks

This thesis shows some of the challenges facing microalgal cell factories, especially those of functionally annotating proteins. While I have made steps on this topic, there are still many improvements required to be made on a wide array of challenges if microalgal biotechnology research is to come of age.

Summary

Chapter 1 provides the background for the thesis. In it, I provide a short overview of what biotechnology is and how it has been utilized it for thousands of years. I then address how modern biotechnology has evolved over the past few decades. Its progress has been triggered by the discovery of nucleic acids and marked by a focus on genetic understanding of cell and organism function and on the subsequent manipulation to ultimately benefit society in one way or another. Furthermore, computational biology has been increasingly important to determine the success of biotechnological research, in for example an anti-malarial drug- producing yeast. However, for microalgae, which are very promising organisms for biotechnological applications, there are essentially no successful commercialized examples of modern biotechnology. The chapter further discusses the importance of computationally predicting protein functions and its role in bioinformatics and systems biology research, concluding that this is one of the challenges for microalgal biotechnology. This topic is discussed in all chapters of this thesis, as its overarching goal is to develop and deploy tools and methodologies that lead to increase our understanding of microalgae as cell factories.

Chapter 2 is a review on the state of microalgal biotechnology in 2014, of which the major discussion points are still valid, and how bioinformatics and systems biology should be used to further microalgal research. It describes the challenges of microalgal genomics, bioinformatics, and systems biology research. The chapter addresses a few challenges for microalgae in particular: a lack of genomic data, a low amount of validated protein functions, and genome-scale metabolic models largely based off of *Arabidopsis thaliana*. Suggestions are made on how to overcome these challenges, by for example better utilizing bioinformatics methods and databases. **Chapter 3** addresses a specific challenge: the need for accurate annotation of the functions of microalgal proteins. It exposes the lack of understanding we have of their protein functions, with a staggering 90% of their annotations also present in the distantly related plant *Arabidopsis thaliana*. Finally, this chapter outlines areas in which microalgal protein function prediction can be improved. In **Chapter 4**, I present CrowdGO, a prediction tool based on the "wisdom of the crowd" principle for protein function prediction that aims to overcome the major problem highlighted in Chapter 3. It operates by taking and merging the existing predictions made by other methods. These merged predictions are then put through a machine learning algorithm which is trained to recognize

patterns in these predictions and correlate them to true or false positives. CrowdGO shows significantly higher accuracy, with a p-value < 2.22e-16, over existing prediction methods, as well as an improved precision and recall optimum.

In **Chapter 5** deploy CrowdGO to the genomics of the oleaginous yeast *Cutaneotrichosporon curvatus*, which thus serves as a real biological test case for the method. Comparisons between the CrowdGO annotated *C. curvatus* proteins to the existing ones of a related yeast showcases the potential of CrowdGO. GO enrichment analysis of *C. curvatus* between transcriptomes of normal growth conditions and nitrogen starved conditions shows cell maintenance functions enriched during the first, and stress functions enriched during the latter. This is in line with what one would expect for an oleaginous eukaryote and reassures us that the CrowdGO annotations are reliable. The CrowdGO annotations are further used in a manual annotation pipeline, which we used to manually curate over 700 metabolic *C. curvatus* proteins. These are used together with differential expression analysis to characterize triacylglycerol synthesis during nitrogen starvation conditions. Only one enzyme was missing after the first round of annotations, displaying a high recall for enzymes when using the manual annotation pipeline.

In **chapter 6** we study the comparative genomics between different *Botryococcus braunii* strains, an oleaginous eukaryote that either makes large number of polysaccharides or hydrocarbons based on the strain. In this chapter, all methodologies discussed or developed in the previous chapters are used to try and identify the key genetic differences between the two strains that lead to polysaccharide or hydrocarbon synthesis. We use CrowdGO to annotate all the proteins and perform manual annotation on a thousand metabolic proteins. These are used in conjunction with quantitative proteomics analysis of several conditions including light and dark, different nitrogen levels, and different cell phases. By combining the manual annotations and the proteomics analysis, we were able to characterize several key pathways including the non-mevalonate pathway, fucose synthesis pathway, and the TCA cycle. Analysis of these pathways reveals key differences in the expression of enzymes that are likely to correspond to polysaccharide or hydrocarbon synthesis. Apart from revealing some key features about *Botryococcus braunii*, this chapter serves as a template for future large-scale microalgal research.

Chapter 7 is a general discussion on the thesis. In it, I discuss how the work in this thesis relates to the SPLASH project for microalgae. Furthermore, I discuss how microalgal annotations can still be improved through the use of various stages of bioinformatics, systems biology, and synthetic / metabolic engineering research. Finally, I discuss how microalgae have potential as protein farms, and how it might be possible to unlock this potential.

Bibliography

1. Lenka, S.K., et al., *Current advances in molecular, biochemical, and computational modeling analysis of microalgal triacylglycerol biosynthesis.* Biotechnology advances, 2016. **34**(5): p. 1046-1063.
2. Falda, M., et al., *Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms.* BMC Bioinformatics, 2012. **13**(Suppl 4): p. S14.
3. Bud, R., *The uses of life: a history of biotechnology*. 1994: Cambridge University Press.
4. Larson, G. and D.Q. Fuller, *The Evolution of Animal Domestication.* Annual Review of Ecology, Evolution, and Systematics, 2014. **45**(1): p. 115-136.
5. LEGRAS, J.L., et al., *Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history.* Molecular ecology, 2007. **16**(10): p. 2091-2102.
6. Development, T.O.f.E.C.-o.a., *Modern biotechnology and the OECD*. 1999.
7. Yanagisawa, S., et al., *Metabolic engineering with Dof1 transcription factor in plants: Improved nitrogen assimilation and growth under low-nitrogen conditions.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(20): p. 7833-7838.
8. Paddon, C.J., et al., *High-level semi-synthetic production of the potent antimalarial artemisinin.* Nature, 2013. **496**: p. 528.
9. Qiao, K., et al., *Engineering lipid overproduction in the oleaginous yeast Yarrowia lipolytica.* Metabolic Engineering, 2015. **29**: p. 56-65.
10. Ye, X., et al., *Engineering the Provitamin A (β-Carotene) Biosynthetic Pathway into (Carotenoid-Free) Rice Endosperm.* Science, 2000. **287**(5451): p. 303-305.
11. Weitzel, M., et al., *13CFLUX2—high-performance software suite for 13C-metabolic flux analysis.* Bioinformatics, 2012. **29**(1): p. 143-145.

12. Bansal, A.K., *Bioinformatics in microbial biotechnology – a mini review.* Microbial Cell Factories, 2005. **4**(1): p. 19.
13. Tripathi, K., *Bioinformatics: The foundation of present and future biotechnology.* Current Science, 2000. **79**(5): p. 570-575.
14. Okafor, N. and B.C. Okeke, *Modern industrial microbiology and biotechnology*. 2017: CRC Press.
15. Yizhak, K., et al., *Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model.* Bioinformatics, 2010. **26**(12): p. i255-i260.
16. Nielsen, J. and J.D. Keasling, *Engineering cellular metabolism.* Cell, 2016. **164**(6): p. 1185-1197.
17. Paddon, C.J. and J.D. Keasling, *Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development.* Nature Reviews Microbiology, 2014. **12**(5): p. 355.
18. Nielsen, J., et al., *Engineering synergy in biotechnology.* Nature chemical biology, 2014. **10**(5): p. 319.
19. Specht, E.A., et al., *Host Organisms: Algae.* Industrial Biotechnology: Microorganisms, 2017: p. 605-641.
20. al., K.G.e., *Algae: An Alternative to the Higher Plant System in Gene Farming.* 2012.
21. Specht, E., S. Miyake-Stoner, and S. Mayfield, *Micro-algae come of age as a platform for recombinant protein production.* Biotechnology Letters, 2010. **32**(10): p. 1373-1383.
22. Lu, J., C. Sheahan, and P. Fu, *Metabolic engineering of algae for fourth generation biofuels production.* Energy & Environmental Science, 2011. **4**(7): p. 2451-2466.
23. Adarme-Vega, T.C., et al., *Microalgal biofactories: a promising approach towards sustainable omega-3 fatty acid production.* Microb Cell Fact, 2012. **11**: p. 96.
24. Jones, C.S. and S.P. Mayfield, *Algae biofuels: versatility for the future of bioenergy.* Curr Opin Biotechnol, 2012. **23**(3): p. 346-51.
25. Lohr, M., J. Schwender, and J.E. Polle, *Isoprenoid biosynthesis in eukaryotic phototrophs: a spotlight on algae.* Plant Sci, 2012. **185-186**: p. 9-22.

26. Lewin, R.A., *Extracellular polysaccharides of green algae.* Canadian Journal of Microbiology, 1956. **2**(7): p. 665-672.
27. Thomas, N.V. and S.K. Kim, *Beneficial effects of marine algal compounds in cosmeceuticals.* Mar Drugs, 2013. **11**(1): p. 146-64.
28. Hu, Q., et al., *Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances.* The plant journal, 2008. **54**(4): p. 621-639.
29. Breuer, G., et al., *The impact of nitrogen starvation on the dynamics of triacylglycerol accumulation in nine microalgae strains.* Bioresource Technology, 2012. **124**: p. 217-226.
30. Feofilova, E.P., E. Sergeeva Ia, and A.A. Ivashechkin, *[Biodiesel-fuel: content, production, producers, contemporary biotechnology (review)].* Prikl Biokhim Mikrobiol, 2010. **46**(4): p. 405-15.
31. Greenwell, H.C., et al., *Placing microalgae on the biofuels priority list: a review of the technological challenges.* J R Soc Interface, 2010. **7**(46): p. 703-26.
32. Sivakumar, G., et al., *Integrated green algal technology for bioremediation and biofuel.* Bioresour Technol, 2012. **107**: p. 1-9.
33. Arad, S.M. and A. Yaron, *Natural pigments from red microalgae for use in foods and cosmetics.* Trends in Food Science & Technology, 1992. **3**: p. 92-97.
34. Spolaore, P., et al., *Commercial applications of microalgae.* J Biosci Bioeng, 2006. **101**(2): p. 87-96.
35. Becker, E., *Micro-algae as a source of protein.* Biotechnology advances, 2007. **25**(2): p. 207-210.
36. Lenihan-Geels, G., K.S. Bishop, and L.R. Ferguson, *Alternative sources of omega-3 fats: can we find a sustainable substitute for fish?* Nutrients, 2013. **5**(4): p. 1301-15.
37. Kim, S.K., T.S. Vo, and D.H. Ngo, *Antiallergic benefit of marine algae in medicinal foods.* Adv Food Nutr Res, 2011. **64**: p. 267-75.

38. Kim, S.K., T.S. Vo, and D.H. Ngo, *Potential application of marine algae as antiviral agents in medicinal foods.* Adv Food Nutr Res, 2011. **64**: p. 245-54.
39. Markou, G. and E. Nerantzis, *Microalgae for high-value compounds and biofuels production: A review with focus on cultivation under stress conditions.* Biotechnol Adv, 2013.
40. Maeda, Y., et al., *Marine microalgae for production of biofuels and chemicals.* Current opinion in biotechnology, 2018. **50**: p. 111-120.
41. Jeon, S., et al., *Current status and perspectives of genome editing technology for microalgae.* Biotechnology for Biofuels, 2017. **10**: p. 267.
42. Chisti, Y., *Constraints to commercialization of algal fuels.* J Biotechnol, 2013. **167**(3): p. 201-14.
43. Day, J.G., S.P. Slocombe, and M.S. Stanley, *Overcoming biological constraints to enable the exploitation of microalgae for biofuels.* Bioresour Technol, 2012. **109**: p. 245-51.
44. Wijffels, R.H. and M.J. Barbosa, *An outlook on microalgal biofuels.* Science, 2010. **329**(5993): p. 796-799.
45. Wijffels, R.H., O. Kruse, and K.J. Hellingwerf, *Potential of industrial biotechnology with cyanobacteria and eukaryotic microalgae.* Curr Opin Biotechnol, 2013. **24**(3): p. 405-13.
46. Wobbe, L. and C. Remacle, *Improving the sunlight-to-biomass conversion efficiency in microalgal biofactories.* Journal of Biotechnology, 2015. **201**: p. 28-42.
47. Sturme, M.H.J., et al., *Transcriptome analysis reveals the genetic foundation for the dynamics of starch and lipid production in Ettlia oleoabundans.* Algal Research, 2018. **33**: p. 142-155.
48. Hagemann, M. and H. Bauwe, *Photorespiration and the potential to improve photosynthesis.* Current Opinion in Chemical Biology, 2016. **35**: p. 109-116.
49. Bar-Even, A., *Daring metabolic designs for enhanced plant carbon fixation.* Plant Science, 2017.
50. Goffeau, A., et al., *Life with 6000 genes.* Science, 1996. **274**(5287): p. 546-567.

51. Blattner, F.R., et al., *The complete genome sequence of Escherichia coli K-12.* science, 1997. **277**(5331): p. 1453-1462.
52. Initiative, A.G., *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* nature, 2000. **408**(6814): p. 796.
53. Derelle, E., et al., *Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features.* Proceedings of the National Academy of Sciences, 2006. **103**(31): p. 11647-11652.
54. Merchant, S.S., et al., *The Chlamydomonas genome reveals the evolution of key animal and plant functions.* Science, 2007. **318**(5848): p. 245-50.
55. Consortium, U., *UniProt: the universal protein knowledgebase.* Nucleic acids research, 2016. **45**(D1): p. D158-D169.
56. Radivojac, P., et al., *A large-scale evaluation of computational protein function prediction.* Nat Methods, 2013. **10**(3): p. 221-7.
57. O'Brien, E.J. and B.O. Palsson, *Computing the functional proteome: recent progress and future prospects for genome-scale models.* Current opinion in biotechnology, 2015. **34**: p. 125-134.
58. Pearson, W.R., *[5] rapid and sensitive sequence comparison with fastp and fasta.* 1990.
59. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.
60. Camacho, C., et al., *BLAST+: architecture and applications.* BMC bioinformatics, 2009. **10**(1): p. 421.
61. Jones, P., et al., *InterProScan 5: genome-scale protein function classification.* Bioinformatics, 2014. **30**(9): p. 1236-1240.
62. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-763.
63. Finn, R.D., et al., *InterPro in 2017—beyond protein family and domain annotations.* Nucleic acids research, 2016. **45**(D1): p. D190-D199.

64. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25.
65. Huntley, R.P., et al., *The GOA database: gene ontology annotation updates for 2015.* Nucleic acids research, 2015. **43**(D1): p. D1057-D1063.
66. Karpowicz, S.J., et al., *The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage.* Journal of Biological Chemistry, 2011. **286**(24): p. 21427-21439.
67. Bernardes, J.S. and C.E. Pedreira, *A review of protein function prediction under machine learning perspective.* Recent Pat Biotechnol, 2013. **7**(2): p. 122-41.
68. Zhu, F., et al., *Homology-free prediction of functional class of proteins and peptides by support vector machines.* Curr Protein Pept Sci, 2008. **9**(1): p. 70-95.
69. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.
70. Joshi-Tope, G., et al., *Reactome: a knowledgebase of biological pathways.* Nucleic acids research, 2005. **33**(suppl_1): p. D428-D432.
71. Kutmon, M., et al., *WikiPathways: capturing the full diversity of pathway knowledge.* Nucleic acids research, 2015. **44**(D1): p. D488-D494.
72. Caspi, R., et al., *The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic acids research, 2007. **36**(suppl_1): p. D623-D631.
73. Dorrell, R.G. and A.G. Smith, *Do red and green make brown? perspectives on plastid acquisitions within the chromalveolates.* Eukaryotic Cell, 2011: p. EC-00326.
74. Borowitzka, M.A., *High-value products from microalgae—their development and commercialisation.* Journal of applied phycology, 2013. **25**(3): p. 743-756.
75. Scott, S.A., et al., *Biodiesel from algae: challenges and prospects.* Curr Opin Biotechnol, 2010. **21**(3): p. 277-86.
76. Wijffels, R.H., M.J. Barbosa, and M.H.M. Eppink, *Microalgae for the production of bulk chemicals and biofuels.* Biofuels,

Bioproducts and Biorefining: Innovation for a sustainable economy, 2010. **4**(3): p. 287-295.
77. Merchant, S.S., et al., *TAG, you're it! Chlamydomonas as a reference organism for understanding algal triacylglycerol accumulation.* Curr Opin Biotechnol, 2012. **23**(3): p. 352-63.
78. Mata, T.M., A.A. Martins, and N.S. Caetano, *Microalgae for biodiesel production and other applications: a review.* Renewable and sustainable energy reviews, 2010. **14**(1): p. 217-232.
79. Passell, H., et al., *Algae biodiesel life cycle assessment using current commercial data.* Journal of environmental management, 2013. **129**: p. 103-111.
80. Jorquera, O., et al., *Comparative energy life-cycle analyses of microalgal biomass production in open ponds and photobioreactors.* Bioresource technology, 2010. **101**(4): p. 1406-1413.
81. Klok, A.J., et al., *Simultaneous growth and neutral lipid accumulation in microalgae.* Bioresource Technology, 2013. **134**: p. 233-243.
82. Busi, M.V., et al., *Starch metabolism in green algae.* Starch-Stärke, 2014. **66**(1-2): p. 28-40.
83. Ho, S.-H., et al., *Bioethanol production using carbohydrate-rich microalgae biomass as feedstock.* Bioresource technology, 2013. **135**: p. 191-198.
84. Hildebrand, M., et al., *Metabolic and cellular organization in evolutionarily diverse microalgae as related to biofuels production.* Current opinion in chemical biology, 2013. **17**(3): p. 506-514.
85. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic acids research, 1997. **25**(17): p. 3389-3402.
86. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic acids research, 2011. **39**(suppl_2): p. W29-W37.
87. Kourmpetis, Y.A., et al., *Bayesian Markov Random Field analysis for protein function prediction based on network data.* PLoS One, 2010. **5**(2): p. e9293.

88. Engelhardt, B.E., et al., *Genome-scale phylogenetic function annotation of large and diverse protein families.* Genome Res, 2011. **21**(11): p. 1969-80.
89. Cozzetto, D., et al. *Protein function prediction by massive integration of evolutionary analyses and multiple data sources*. BioMed Central.
90. Wong, A. and H. Shatkay, *Protein Function Prediction using Text-based Features extracted from the Biomedical Literature: The CAFA Challenge.* BMC Bioinformatics, 2013. **14**(Suppl 3): p. S14.
91. Buchan, D.W., et al., *Protein annotation and modelling servers at University College London.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W563-8.
92. Rentzsch, R. and C.A. Orengo, *Protein function prediction--the power of multiplicity.* Trends Biotechnol, 2009. **27**(4): p. 210-9.
93. Rodgers-Melnick, E., M. Culp, and S.P. Difazio, *Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS.* BMC Genomics, 2013. **14**(1): p. 608.
94. Franceschini, A., et al., *STRING v9.1: protein-protein interaction networks, with increased coverage and integration.* Nucleic Acids Res, 2013. **41**(Database issue): p. D808-15.
95. Emanuelsson, O., et al., *Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence.* Journal of Molecular Biology, 2000. **300**(4): p. 1005-1016.
96. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nature methods, 2011. **8**(10): p. 785.
97. Nancy, Y.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.* Bioinformatics, 2010. **26**(13): p. 1608-1615.
98. Tardif, M., et al., *PredAlgo: a new subcellular localization prediction tool dedicated to green algae.* Molecular biology and evolution, 2012: p. mss178.

99. Wienkoop, S., et al., *Targeted proteomics for Chlamydomonas reinhardtii combined with rapid subcellular protein fractionation, metabolomics and metabolic flux analyses.* Molecular BioSystems, 2010. **6**(6): p. 1018-1031.
100. Terashima, M., M. Specht, and M. Hippler, *The chloroplast proteome: a survey from the Chlamydomonas reinhardtii perspective with a focus on distinctive features.* Current genetics, 2011. **57**(3): p. 151-168.
101. May, P., et al., *ChlamyCyc: an integrative systems biology database and web-portal for Chlamydomonas reinhardtii.* Bmc Genomics, 2009. **10**(1): p. 209.
102. Wienkoop, S., et al., *ProMEX–a mass spectral reference database for plant proteomics.* Frontiers in plant science, 2012. **3**: p. 125.
103. Keller, O., et al., *A novel hybrid gene prediction method employing protein multiple sequence alignments.* Bioinformatics, 2011. **27**(6): p. 757-763.
104. Lopez, D., et al., *Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data.* Bmc Bioinformatics, 2011. **12**(1): p. 282.
105. Goodstein, D.M., et al., *Phytozome: a comparative platform for green plant genomics.* Nucleic acids research, 2011. **40**(D1): p. D1178-D1186.
106. Bogen, C., et al., *Reconstruction of the lipid metabolism for the microalga Monoraphidium neglectum from its genome sequence reveals characteristics suitable for biofuel production.* BMC genomics, 2013. **14**(1): p. 926.
107. Desbois, A.P., et al., *Isolation and structural characterisation of two antibacterial free fatty acids from the marine diatom, Phaeodactylum tricornutum.* Applied microbiology and biotechnology, 2008. **81**(4): p. 755-764.
108. Rosenberg, J.N., et al., *Comparative analyses of three Chlorella species in response to light and sugar reveal distinctive lipid accumulation patterns in the microalga C. sorokiniana.* PloS one, 2014. **9**(4): p. e92460.

109. Gao, C., et al., *Oil accumulation mechanisms of the oleaginous microalga Chlorella protothecoides revealed through its genome, transcriptomes, and proteomes.* BMC Genomics, 2014. **15**: p. 582.
110. Doucha, J. and K. Lívanský, *Production of high-density Chlorella culture grown in fermenters.* Journal of applied phycology, 2012. **24**(1): p. 35-43.
111. Feng, Y., C. Li, and D. Zhang, *Lipid production of Chlorella vulgaris cultured in artificial wastewater medium.* Bioresource technology, 2011. **102**(1): p. 101-105.
112. Prieto, A., J.P. Canavate, and M. García-González, *Assessment of carotenoid production by Dunaliella salina in different culture systems and operation regimes.* Journal of biotechnology, 2011. **151**(2): p. 180-185.
113. Ambati, R.R., et al., *Astaxanthin: sources, extraction, stability, biological activities and its commercial applications—a review.* Marine drugs, 2014. **12**(1): p. 128-152.
114. Weiss, T.L., et al., *Genome size and phylogenetic analysis of the A and L races of Botryococcus braunii.* Journal of applied phycology, 2011. **23**(5): p. 833-839.
115. Talukdar, J., M.C. Kalita, and B.C. Goswami, *Characterization of the biofuel potential of a newly isolated strain of the microalga Botryococcus braunii Kützing from Assam, India.* Bioresource technology, 2013. **149**: p. 268-275.
116. Weiss, T.L., et al., *Colony organization in the green alga Botryococcus braunii (Race B) is specified by a complex extracellular matrix.* Eukaryotic cell, 2012. **11**(12): p. 1424-1440.
117. Orth, J.D., I. Thiele, and B.Ø. Palsson, *What is flux balance analysis?* Nature biotechnology, 2010. **28**(3): p. 245.
118. Mahadevan, R. and C.H. Schilling, *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metabolic engineering, 2003. **5**(4): p. 264-276.

119. Boyle, N.R. and J.A. Morgan, *Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii.* BMC systems biology, 2009. **3**(1): p. 4.
120. Zomorrodi, A.R., et al., *Mathematical optimization applications in metabolic networks.* Metabolic engineering, 2012. **14**(6): p. 672-686.
121. May, P., et al., *Metabolomics-and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii.* Genetics, 2008. **179**(1): p. 157-166.
122. Christian, N., et al., *An integrative approach towards completing genome-scale metabolic networks.* Molecular BioSystems, 2009. **5**(12): p. 1889-1903.
123. Manichaikul, A., et al., *Metabolic network analysis integrated with transcript verification for sequenced genomes.* Nature methods, 2009. **6**(8): p. 589.
124. Chang, R.L., et al., *Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism.* Mol Syst Biol, 2011. **7**: p. 518.
125. Krumholz, E.W., et al., *Genome-wide metabolic network reconstruction of the picoalga Ostreococcus.* Journal of experimental botany, 2011. **63**(6): p. 2353-2362.
126. Kliphuis, A.M.J., et al., *Metabolic modeling of Chlamydomonas reinhardtii: energy requirements for photoautotrophic growth and maintenance.* Journal of applied phycology, 2012. **24**(2): p. 253-266.
127. Cogne, G., et al., *A model-based method for investigating bioenergetic processes in autotrophically growing eukaryotic microalgae: Application to the green algae Chlamydomonas reinhardtii.* Biotechnology progress, 2011. **27**(3): p. 631-640.
128. Rivas, M.O., P. Vargas, and C.E. Riquelme, *Interactions of Botryococcus braunii cultures with bacterial biofilms.* Microbial ecology, 2010. **60**(3): p. 628-635.
129. Zomorrodi, A.R. and C.D. Maranas, *OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities.* PLoS computational biology, 2012. **8**(2): p. e1002363.

130. Zomorrodi, A.R., M.M. Islam, and C.D. Maranas, *d-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities.* ACS synthetic biology, 2014. **3**(4): p. 247-257.
131. Tomar, N. and R.K. De, *Comparing methods for metabolic network analysis and an application to metabolic engineering.* Gene, 2013. **521**(1): p. 1-14.
132. Cho, A., et al., *Prediction of novel synthetic pathways for the production of desired chemicals.* BMC Systems Biology, 2010. **4**(1): p. 35.
133. Wiechert, W., *13C metabolic flux analysis.* Metabolic engineering, 2001. **3**(3): p. 195-206.
134. Jensen, P.A., K.A. Lutz, and J.A. Papin, *TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks.* BMC systems biology, 2011. **5**(1): p. 147.
135. Blazier, A.S. and J.A. Papin, *Integration of expression data in genome-scale metabolic network reconstructions.* Frontiers in physiology, 2012. **3**: p. 299.
136. Lehr, F. and C. Posten, *Closed photo-bioreactors as tools for biofuel production.* Current opinion in biotechnology, 2009. **20**(3): p. 280-285.
137. Kirst, H., et al., *Truncated photosystem chlorophyll antenna size in the green microalga Chlamydomonas reinhardtii upon deletion of the TLA3-CpSRP43 gene.* Plant physiology, 2012: p. pp-112.
138. Bar-Even, A., et al., *Design and analysis of synthetic carbon fixation pathways.* Proceedings of the National Academy of Sciences, 2010. **107**(19): p. 8889-8894.
139. Whitney, S.M., R.L. Houtz, and H. Alonso, *Advancing our understanding and capacity to engineer nature's CO2-sequestering enzyme, Rubisco.* Plant Physiol, 2011. **155**(1): p. 27-35.
140. Guarnieri, M.T., et al., *Proteomic analysis of Chlorella vulgaris: potential targets for enhanced lipid accumulation.* Journal of proteomics, 2013. **93**: p. 245-253.
141. Trentacoste, E.M., et al., *Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without*

compromising growth. Proceedings of the National Academy of Sciences, 2013. **110**(49): p. 19748-19753.
142. Dahl, R.H., et al., *Engineering dynamic pathway regulation using stress-response promoters.* Nature biotechnology, 2013. **31**(11): p. 1039.
143. Kilian, O., et al., *High-efficiency homologous recombination in the oil-producing alga Nannochloropsis sp.* Proceedings of the National Academy of Sciences, 2011. **108**(52): p. 21265-21269.
144. Purton, S., et al., *Genetic engineering of algal chloroplasts: progress and prospects.* Russian journal of plant physiology, 2013. **60**(4): p. 491-499.
145. Ohresser, M., R.F. Matagne, and R. Loppes, *Expression of the arylsulphatase reporter gene under the control of the nit1 promoter in Chlamydomonas reinhardtii.* Current genetics, 1997. **31**(3): p. 264-271.
146. Quinn, J.M., J. Kropat, and S. Merchant, *Copper response element and Crr1-dependent Ni2+-responsive promoter for induced, reversible gene expression in Chlamydomonas reinhardtii.* Eukaryotic Cell, 2003. **2**(5): p. 995-1002.
147. Helliwell, K.E., et al., *Unravelling vitamin B12-responsive gene regulation in algae.* Plant physiology, 2014: p. pp-113.
148. Ramundo, S., et al., *Repression of essential chloroplast genes reveals new signaling pathways and regulatory feedback loops in Chlamydomonas.* The Plant Cell, 2013: p. tpc-112.
149. Fang, W., et al., *Transcriptome-wide changes in Chlamydomonas reinhardtii gene expression regulated by carbon dioxide and the CO2-concentrating mechanism regulator CIA5/CCM1.* The Plant Cell, 2012: p. tpc-112.
150. Rasala, B.A., et al., *Robust expression and secretion of Xylanase1 in Chlamydomonas reinhardtii by fusion to a selection gene and processing with the FMDV 2A peptide.* PloS one, 2012. **7**(8): p. e43349.
151. Day, A. and M. Goldschmidt-Clermont, *The chloroplast transformation toolbox: selectable markers and marker removal.* Plant biotechnology journal, 2011. **9**(5): p. 540-553.

152. Bordbar, A., et al., *Constraint-based models predict metabolic and associated cellular functions.* Nat Rev Genet, 2014. **15**(2): p. 107-120.
153. Leliaert, F., et al., *Phylogeny and Molecular Evolution of the Green Algae.* Critical Reviews in Plant Sciences, 2012. **31**(1): p. 1-46.
154. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**(1): p. 421.
155. Harris, E.H., *CHLAMYDOMONAS AS A MODEL ORGANISM.* Annu Rev Plant Physiol Plant Mol Biol, 2001. **52**: p. 363-406.
156. Scaife, M.A., et al., *Establishing Chlamydomonas reinhardtii as an industrial biotechnology host.* The Plant Journal, 2015. **82**(3): p. 532-546.
157. Reijnders, M.J., et al., *Green genes: bioinformatics and systems-biology innovations drive algal biotechnology.* Trends in biotechnology, 2014. **32**(12): p. 617-626.
158. Grossman, A.R., et al., *Chlamydomonas reinhardtii at the crossroads of genomics.* Eukaryot Cell, 2003. **2**(6): p. 1137-50.
159. Consortium, U., *UniProt: a hub for protein information.* Nucleic acids research, 2014: p. gku989.
160. Moreau, H., et al., *Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage.* Genome Biol, 2012. **13**(8): p. R74.
161. Blanc, G., et al., *The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex.* Plant Cell, 2010. **22**(9): p. 2943-55.
162. Blanc, G., et al., *The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation.* Genome Biol, 2012. **13**(5): p. R39.
163. Worden, A.Z., et al., *Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas.* Science, 2009. **324**(5924): p. 268-72.
164. Blanc-Mathieu, R., et al., *An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing*

Illumina de novo assemblies. BMC Genomics, 2014. **15**: p. 1103.
165. Palenik, B., et al., *The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation.* Proc Natl Acad Sci U S A, 2007. **104**(18): p. 7705-10.
166. Prochnik, S.E., et al., *Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri.* Science, 2010. **329**(5988): p. 223-6.
167. Minneci, F., et al., *FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences.* 2013.
168. Suzek, B.E., et al., *UniRef: comprehensive and non-redundant UniProt reference clusters.* Bioinformatics, 2007. **23**(10): p. 1282-1288.
169. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic acids research, 2013: p. gkt1223.
170. Biasini, M., et al., *SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information.* Nucleic acids research, 2014: p. gku340.
171. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat. Protocols, 2010. **5**(1): p. 93-121.
172. Devoid, S., et al., *Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED*, in *Systems Metabolic Engineering*. 2013, Springer. p. 17-45.
173. Karp, P.D., et al., *Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.* Briefings in bioinformatics, 2009: p. bbp043.
174. Millar, A.H. and N.L. Taylor, *Subcellular proteomics—where cell biology meets protein chemistry.* Frontiers in Plant Science, 2014. **5**: p. 55.
175. Barsanti, L. and P. Gualtieri, *Algae: anatomy, biochemistry, and biotechnology*. 2014: CRC press.
176. Chibucos, M.C., et al., *Standardized description of scientific evidence using the Evidence Ontology (ECO).* Database, 2014. **2014**.

177. Schaid, D.J., et al., *Using the Gene Ontology to Scan Multi-Level Gene Sets for Associations in Genome Wide Association Studies.* Genetic epidemiology, 2012. **36**(1): p. 3-16.
178. Conesa, A., et al., *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.* Bioinformatics, 2005. **21**(18): p. 3674-3676.
179. Worm, P., et al., *A genomic view on syntrophic versus non-syntrophic lifestyle in anaerobic fatty acid degrading communities.* Biochimica et Biophysica Acta (BBA) - Bioenergetics, 2014. **1837**(12): p. 2004-2016.
180. Elliott, L.G., et al., *Establishment of a bioenergy-focused microalgal culture collection.* Algal Research, 2012. **1**(2): p. 102-113.
181. Lan, L., et al. *MS-k NN: protein function prediction by integrating multiple data sources*. in *BMC bioinformatics*. 2013. BioMed Central.
182. Jiang, Y., et al., *An expanded evaluation of protein function prediction methods shows an improvement in accuracy.* Genome biology, 2016. **17**(1): p. 184.
183. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
184. Hearst, M.A., et al., *Support vector machines.* IEEE Intelligent Systems and their applications, 1998. **13**(4): p. 18-28.
185. Finn, R.D., et al., *HMMER web server: 2015 update.* Nucleic acids research, 2015. **43**(W1): p. W30-W38.
186. Consortium, G.O., *Expansion of the Gene Ontology knowledgebase and resources.* Nucleic acids research, 2016. **45**(D1): p. D331-D338.
187. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC bioinformatics, 2011. **12**(1): p. 77.
188. Lin, D. *An information-theoretic definition of similarity*. in *Icml*. 1998. Citeseer.
189. Grau, J., I. Grosse, and J. Keilwagen, *PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R.* Bioinformatics, 2015. **31**(15): p. 2595-2597.

190. Johnson, E.A., *Biotechnology of non-Saccharomyces yeasts—the ascomycetes.* Applied microbiology and biotechnology, 2013. **97**(2): p. 503-517.
191. Meesters, P.A.E.P., G.N.M. Huijberts, and G. Eggink, *High-cell-density cultivation of the lipid accumulating yeast Cryptococcus curvatus using glycerol as a carbon source.* Applied Microbiology and Biotechnology, 1996. **45**(5): p. 575-579.
192. Bednarski, W., J. Leman, and J. Tomasik, *Utilization of beet molasses and whey for fat biosynthesis by a yeast.* Agricultural Wastes, 1986. **18**(1): p. 19-26.
193. Chang, Y.-H., et al., *Microbial lipid production by oleaginous yeast Cryptococcus sp. in the batch cultures using corncob hydrolysate as carbon source.* biomass and bioenergy, 2015. **72**: p. 95-103.
194. Reijnders, M.J., B.M. Carreres, and P.J. Schaap, *Algal omics: The functional annotation challenge.* Current Biotechnology, 2015. **4**(4): p. 457-463.
195. Taylor, J.W. and M.L. Berbee, *Dating divergences in the Fungal Tree of Life: review and new analyses.* Mycologia, 2006. **98**(6): p. 838-849.
196. Ykema, A., et al., *Optimization of lipid production in the oleaginous yeastApiotrichum curvatum in wheypermeate.* Applied microbiology and biotechnology, 1988. **29**(2-3): p. 211-218.
197. Odoni, D.I., et al., *Aspergillus niger Secretes Citrate to Increase Iron Bioavailability.* Frontiers in microbiology, 2017. **8**: p. 1424.
198. Loira, N., et al., *A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica.* BMC Systems Biology, 2012. **6**(1): p. 35.
199. McWilliam, H., et al., *Analysis tool web services from the EMBL-EBI.* Nucleic acids research, 2013. **41**(W1): p. W597-W600.
200. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

201. Young, M.D., et al., *goseq: Gene Ontology testing for RNA-seq datasets.* R Bioconductor, 2012.
202. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms.* PloS one, 2011. **6**(7): p. e21800.
203. Close, D. and J. Ojumu, *Draft genome sequence of the oleaginous yeast Cryptococcus curvatus ATCC 20509.* Genome announcements, 2016. **4**(6): p. e01235-16.
204. Banerjee, A., et al., *Botryococcus braunii: a renewable source of hydrocarbons and other chemicals.* Critical reviews in biotechnology, 2002. **22**(3): p. 245-279.
205. Molnar, I., et al., *Bio-crude transcriptomics: gene discovery and metabolic network reconstruction for the biosynthesis of the terpenome of the hydrocarbon oil-producing green alga, Botryococcus braunii race B (Showa).* BMC Genomics, 2012. **13**: p. 576.
206. Li, Y., et al., *Extractable liquid, its energy and hydrocarbon content in the green alga Botryococcus braunii.* biomass and bioenergy, 2013. **52**: p. 103-112.
207. Thapa, H.R., et al., *A squalene synthase-like enzyme initiates production of tetraterpenoid hydrocarbons in Botryococcus braunii Race L.* Nature communications, 2016. **7**: p. 11198.
208. Ishimatsu, A., et al., *Biosynthesis of isoprene units in the C34 botryococcene molecule produced by Botryococcus braunii strain Bot-22.* Procedia Environmental Sciences, 2012. **15**: p. 56-65.
209. Komárek, J. and P. Marvan, *Morphological differences in natural populations of the genus Botryococcus (Chlorophyceae).* Archiv für Protistenkunde, 1992. **141**(1-2): p. 65-100.
210. Plain, N., et al., *Variabilité morphologique de Botryococcus braunii (Chlorococcales, Chlorophyta): corrélations avec les conditions de croissance et la teneur en lipides.* Phycologia, 1993. **32**(4): p. 259-265.
211. Darienko, T., et al., *Evaluating the species boundaries of green microalgae (Coccomyxa, Trebouxiophyceae, Chlorophyta) using integrative taxonomy and DNA barcoding with further implications for the species*

identification in environmental samples. PloS one, 2015. **10**(6): p. e0127838.
212. Kawachi, M., et al., *Relationship between hydrocarbons and molecular phylogeny of Botryococcus braunii.* Algal Research, 2012. **1**(2): p. 114-119.
213. Hegedűs, A., et al., *Molecular phylogeny of Botryococcus braunii strains (race A)–An integrative approach.* Algal Research, 2016. **19**: p. 189-197.
214. Metzger, P., et al., *Alkadiene-and botryococcene-producing races of wild strains of Botryococcus braunii.* Phytochemistry, 1985. **24**(10): p. 2305-2312.
215. Gouveia, J.D., et al., *Botryococcus braunii strains compared for biomass productivity, hydrocarbon and carbohydrate content.* Journal of biotechnology, 2017. **248**: p. 77-86.
216. Largeau, C., et al., *Sites of accumulation and composition of hydrocarbons in Botryococcus braunii.* Phytochemistry, 1980. **19**(6): p. 1043-1051.
217. Dubois, M., et al., *Colorimetric method for determination of sugars and related substances.* Analytical chemistry, 1956. **28**(3): p. 350-356.
218. Lemieux, C., C. Otis, and M. Turmel, *Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae.* BMC evolutionary biology, 2014. **14**(1): p. 211.
219. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC bioinformatics, 2010. **11**(1): p. 119.
220. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic acids research, 2004. **32**(5): p. 1792-1797.
221. Capella-Gutiérrez, S., J.M. Silla-Martínez, and T. Gabaldón, *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.* Bioinformatics, 2009. **25**(15): p. 1972-1973.
222. Smith, S.A. and C.W. Dunn, *Phyutility: a phyloinformatics tool for trees, alignments and molecular data.* Bioinformatics, 2008. **24**(5): p. 715-716.

223. Guindon, S., et al., *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.* Systematic biology, 2010. **59**(3): p. 307-321.
224. Chevenet, F., et al., *TreeDyn: towards dynamic graphics and annotations for analyses of trees.* BMC bioinformatics, 2006. **7**(1): p. 439.
225. Höhner, R., et al., *The metabolic status drives acclimation of iron deficiency responses in Chlamydomonas reinhardtii as revealed by proteomics based hierarchical clustering and reverse genetics.* Molecular & Cellular Proteomics, 2013. **12**(10): p. 2774-2790.
226. Leufken, J., et al., *pyQms enables universal and accurate quantification of mass spectrometry data.* Molecular & Cellular Proteomics, 2017. **16**(10): p. 1736-1745.
227. Barth, J., et al., *The interplay of light and oxygen in the reactive oxygen stress response of Chlamydomonas reinhardtii dissected by quantitative mass spectrometry.* Molecular & Cellular Proteomics, 2014. **13**(4): p. 969-989.
228. Bergner, S.V., et al., *STATE TRANSITION7-dependent phosphorylation is modulated by changing environmental conditions, and its absence triggers remodeling of photosynthetic protein complexes.* Plant physiology, 2015. **168**(2): p. 615-634.
229. Metzger, P., et al., *Structures of some botryococcenes: branched hydrocarbons from the B-race of the green alga Botryococcus braunii.* Phytochemistry, 1985. **24**(12): p. 2995-3002.
230. Programme, E.C.t.F. *Final Report Summary - SPLASH (Sustainable PoLymers from Algae Sugars and Hydrocarbons)*. 2013; Available from: https://cordis.europa.eu/result/rcn/212762_en.html.
231. Koskinen, P., et al., *PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment.* Bioinformatics, 2015. **31**(10): p. 1544-1552.
232. Chitale, M., et al., *ESG: extended similarity group method for automated protein function prediction.* Bioinformatics, 2009. **25**(14): p. 1739-45.

233. Zhu, M., et al., *Predicting gene regulatory networks of soybean nodulation from RNA-Seq transcriptome data.* BMC bioinformatics, 2013. **14**(1): p. 278.
234. Consortium, U. *UniProt expert biocuration*. 2018; Available from: https://www.uniprot.org/help/?fil=section:biocuration.
235. Corteggiani Carpinelli, E., et al., *Chromosome scale genome assembly and transcriptome profiling of Nannochloropsis gaditana in nitrogen depletion.* Mol Plant, 2014. **7**(2): p. 323-35.
236. Devoid, S., et al., *Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED.* Methods Mol Biol, 2013. **985**: p. 17-45.
237. Tevatia, R., Y. Demirel, and P. Blum, *Kinetic modeling of photoautotropic growth and neutral lipid accumulation in terms of ammonium concentration in Chlamydomonas reinhardtii.* Bioresource technology, 2012. **119**: p. 419-424.
238. Lian, J., et al., *The effect of the algal microbiome on industrial production of microalgae.* Microbial Biotechnology, 2018.
239. Guiry, M.D., *How many species of algae are there?* Journal of phycology, 2012. **48**(5): p. 1057-1063.
240. Wang, Q., et al., *Genome editing of model oleaginous microalgae Nannochloropsis spp. by CRISPR/Cas9.* The Plant Journal, 2016. **88**(6): p. 1071-1081.
241. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific data, 2016. **3**.
242. Koehorst, J.J., et al., *Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data.* Scientific reports, 2016. **6**: p. 38699.
243. Koehorst, J., et al., *Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics [version 1; referees: awaiting peer review].* F1000Research, 2016. **5**.
244. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing*

technologies. Nature Reviews Genetics, 2016. **17**(6): p. 333.
245. Waterhouse, R.M., et al., *BUSCO applications from quality assessments to gene prediction and phylogenomics.* Molecular biology and evolution, 2017. **35**(3): p. 543-548.
246. Hoff, K.J., et al., *BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.* Bioinformatics, 2016. **32**(5): p. 767-769.
247. Cantarel, B.L., et al., *MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.* Genome Research, 2008. **18**(1): p. 188-196.
248. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome biology, 2013. **14**(4): p. R36.
249. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357.
250. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
251. Love, M., S. Anders, and W. Huber, *Differential analysis of count data–the DESeq2 package.* Genome Biol, 2014. **15**(550): p. 10-1186.

Acknowledgements

It is finally done! After years of hard work, I have this book to show as a result. Of course, I met many people along the way, and I have many people to thank for a variety of reasons.

First, I would like to thank Dr. Peter Schaap, Prof. Vitor Martins dos Santos, and Prof. Gerrit Eggink. Without them, this thesis would not have a start, middle, and end.

Peter, you were the main reason I got hired. As a clueless 23-year-old I emailed you after I had a few drinks with **Jasper Koehorst** in the Vlaamsche Reus, and you emailed me back with a job advertisement. The description initially scared me due to all the modeling language in it. Luckily you convinced me it wasn't all that scary, and perhaps even better for me, the PhD turned out to be much more bioinformatics oriented than everyone expected. If there is one thing I can contribute to your supervision, it is the attention for detail. I often get stuck in a general overview ignoring the important details, but you were there to point out flaws or opportunities in my research that I did not see. This is something I am still slowly trying to improve, but your guidance has definitely helped.

Vitor, thank you for hiring me even after I got a basic physics question wrong during my job interview when you were explaining the concept of modeling to me. Naturally, I did not discuss my work with you as much as I did with Peter, but every time we did you came with valuable input. Because of the more limited time we had discussing science, you often sticked to giving more general input. For example, I remember that a big part of our successful review chapter 2 got rewritten after your criticism, and we can't argue with the results. If there is one person I would trust to sell science or ideas to an audience of any kind, it would be you. Furthermore, you were the first to open my eyes to the necessity of networking in science. And as you often showed, the best way for this to be done is with some C2H6O.

Gerrit, because you were in a different chair group we did not see each other that often. But luckily you were there when I needed you, because what would I have done without the data you had lying around for us to analyze. The work on *Cryptococcus* is still ongoing and I hope **Nhung** is treating the data well, so she can somewhere in the future finish what we have started years ago. Thanks for all the advice you gave me during my PhD!

Now to all my colleagues in SSB! What a (weird) time we had together. I can honestly say that you guys were the most fun colleagues I ever had, and I doubt I will enjoy such a good atmosphere anywhere else in the future. I am not sure if that is because the rest of the scientific world is boring, or if I'm just getting older and more mature (haha…). The first person I want to address is not just because he addressed me first in his own acknowledgements, but also because we started in a two-week timespan from each other in the same office.

Ruben, what can I say. I came into SSB knowing some of the master thesis students, and they dragged me for a 'coffee break' once or twice a day. This break basically meant us walking to the coffee machine, getting coffee or tea, and walking back to the office right away. Luckily, you set me straight real quick. "I really don't get it!", is what I remember you saying. As everyone knows, the rest is history. And the coffee breaks in SSB were always amazing, so I thank you for that. We did many things together during our time at SSB, like the VLAG week where we drank so much all the VLAG weeks after us got a limit on their alcohol consumption (true story, right Benoit, Kees, Tijn?). Sadly, the timeline I made there for my noble prizes is a little bit off, but I still got a better score than you for my presentation ;). On a more serious note, I also want to thank you for all the times I had questions about modeling or related topics. You were always patient enough to answer. I could not have wished for a better officemate to start my scientific career with than you.

While this person should probably be way down the list of acknowledgements, I should stick with officemates first. I'm just kidding, **Niru**! You were one of the first people I talked to on the first day of my job, because you walked into the office thinking you would have it all for yourself that day. It must have been so weird for you to suddenly see a homeless looking 24-year-old sitting there, having a big grin on his face. I even managed to be nice to you for at least two weeks! Or was it one? I can't remember… Anyway, THANKS Niru, for being my officemate. You are genuinely a fun person to hang around with, but I'm not sure if it is because you are fun, or if it is because it is easy to make fun of you. We never really helped each other out on a scientific level, but you were always there when I needed for you, despite me being an *******! Proof: you helping me with making sure there will actually be people at my PhD defense after-party. Or is this just some elaborate prank, where you pretend people are actually coming, and in the end no one will show up… It would be the ultimate 'HAH!', but you are too nice for that. Sorry for being weird throughout my entire thesis, but I know you liked my weirdness. I even heard from a little birdy that you missed me making fun of you during coffee breaks.

Of course, I can't leave out **Bart** as one of the first I have to acknowledge. Like with Ruben, we started around the same time. You were more the silent force in our office, with luckily for you a very good headset. However, as our lord of everything data and server related, you could not get around us. I am so sorry for always hogging up the data on SSB3… I swear, it is because I worked with eukaryotes and everyone else with prokaryotes… Anyway, with a few very loud (literally) personalities in SSB it was sometimes hard to notice you. But when you did speak up in a bigger group you usually had some substance to your words, unlike some of the other people in SSB, myself included. In smaller groups however, there was no getting around it: you are a fun and funny guy to hang around with. A bit tech geeky, but not as bad as **Benoit** so that's ok.

I have to talk about **Benoit**… This guy just can't leave me alone! I finally got rid of him after moving to Lausanne, and then he joins me one year later! Luckily, I got away with not seeing him that often, but I'm not sure if I can keep that up forever. Just kidding, of course. I met this guy through our joint friends from my bachelors, who were doing their masters in Wageningen. When I moved to Ede to start my masters thesis in Wageningen, I joined them almost every Thursday for drinks at the Vlaamsche Reus. And there was this weird French guy always drinking Kasteel Donker. We applied for a job at SSB around the same time, and people thought you got my job, which I already thought was mine. They were afraid to tell me, but little did everyone realise we actually both got a job on microalgae at SSB. Sadly, we didn't collaborate that much during our time there, even after moving to an office together. You worked more with BPE, and I worked more with SPLASH. Maybe that is for the best, after going through the process of our review paper together… We did collaborate very well outside of work though: in bars, at barbecues, and in (drunk) WWE fights. You are one weird dude, but you're a fun dude. I am eternally grateful for you helping me move to Switzerland, and for being my paranymph. I am sure that once I am done with my current job in Switzerland, we will have many more stories to tell. Let's never go to that shit club in Lausanne again. And please remember the way back next time…

Tjerko, your face always looked very familiar to me. Like I knew you from before my time at SSB. Only years into my PhD I realized it was probably because you were also a frequent Dollar$ visitor. Sadly, I don't live in Nijmegen anymore. If I met you before I moved out of the city, I'm sure we would have had a great time there. We never hung out too much during the first half of our PhD, mainly because you worked most of your time at MSD. But that changed after we moved into an office together. You are fun, sometimes weird, and have a great sense of humor. Sorry for all the trouble Ben and I caused in the office. I don't think you were too much, as evident by your extremely quick and successful defense of your PhD. You always came across as very well-organized, which might come with

your company background. Something I can learn from! Keep on cycling, and let's hope we meet some time in the future so we can have a drink, work or non-work related.

Carolyn, you are gone for a long time already! But you took me by the hand for the first 1.5 years or so. And I am really grateful for that. We produced an excellent review together, mostly due to your work of tying everything and everyone together! We saw each other a few times during SPLASH meetings, and it was always nice to talk to you again. I hope you are having a great time back in Hong Kong!

Now for mine and everyone's favourite PhD student (totally legit vote by the way), **Bastian**. What can I say… I think you are the only guy in Wageningen that can outweird me without even trying… But you are such a fun guy to hang out with! What would SSB, or Wageningen, be without you?! Not only are you party and nasty basti, you are also responsible Basti! Making sure everyone gets home safe when they drank too much. Although… you never really made sure I got home safe. You were more 'concerned' about the girls… With the exception of course when I left my jacket at some student party, with my keys in it, and I was too drunk to realise. Sorry for waking you up at 4 am in the morning, but thanks for letting me stay the night. Apart from being responsible Basti, I witnessed your transformation from boring Basti to party (and nasty) Basti. They will teach about you in history class. And history will fade into legend. And legend will fade into myth. But we will remember. Ben and I are waiting for you in Lausanne, you still didn't come and visit us and make the city unsafe. Thanks for always answering my questions about science, and for being my paranymph even if it is only two weeks after your own defense. This shows, once more, what a great and kind human being you are. Keep on being you.

Another legend that can't be forgotten is **Rob.** You started shortly after me and have been my go-to post-doc throughout my entire time at SSB. If I had questions about dynamic modeling or mathematics, you were always patient enough to help me out. You

were also a great frontrunner in leading the successful iGEM team of 2016, which I had a great time supervising. Apart from all this boring stuff, I think you are the one guy I hung out the most with in bars during my time in Wageningen. Always in for a drink, or two, or three, until closing time. We had many good discussions, many arguments, but never bad enough to stop hanging out. We're just two passionate, drunk, drunk guys on Fridays. I am genuinely happy for you that you found a job at Oxford, and I hope you will find everything you are looking for back in your home country. As of writing this, I still don't know if you will be able to visit for my PhD defense, but if you are here, we better get wasted.

I am being repetitive so far, but **Nhung**, you are one weird girl. I bet you are drinking an algae smoothie right now. Apart from being a fun person during our time at SSB and being surprisingly party-loving for such an innocent looking girl, I mainly want to thank you for our work on *Cryptococcus*. You took the modeling part out of my hands, THANK GOD. It would have been a disaster if I had to do that! I hope I gave you all the tools I can give you for you to finish the job. The two weeks we spent locked up manually annotating GO terms and EC numbers to proteins were actually a lot of fun, despite it being such a boring and monotonous job. I regret letting you hear the song 'Friday' from Rebecca Black, though… One last request: please stop including me in the authors for your *Cryptococcus* abstract submissions to conferences. I am getting extremely jealous because of the talks and nice trips you are getting! Enjoy the rest of your time in Wageningen, and if you are ever around, be sure to come and visit.

Emma, you disguise yourself as a normal sane person, but you are one of the weirdest of them all. Often too busy or serious to join our coffee breaks, but when you were there, it was often you steering the conversation to questionable topics. When Rob wasn't there, you helped me out with mathematics related questions whenever I needed it. And we made the great bioinformatics and modeling trifecta that led the iGEM team to their success (yeah yeah, the

students also had something to do with it…). The times we had drinks together were always great fun, especially our trip to BBK! Please don't tell anyone about the bird… Good luck with the rest of your PhD, and I am sure you will have success after, be it academics or elsewhere. Like cycling to China and starting an Irish pub.

Jasper, thank you so much for encouraging me that night at the Vlaamsche to send an email to Peter. It got me the job, and if it wasn't for you I wouldn't have met all these amazing people. We go way back, drinking way too many beer$ in Dollar$ every Thursday. You settled and slowed down a bit after that, and I didn't, so the times weren't as crazy as before. But you were always there when I had questions about databases, domains, phylogeny, etc. etc. You are a genuinely hard-working guy, and I think you will stay at SSB forever. How long before you take over Peters job?

Of course, I have to mention **Jesse** right after Jasper. Thanks for the all the questions you answered, and for helping me out with database related queries that needed to be done. You know everything there is to know about the semantic web, and if I will ever be in a database crisis, you will be my go-to guy.

Erika, please don't stare at me creepily when I am defending my thesis… It makes me really uncomfortable. Maybe even a weirder woman than Emma. But thanks for being a very kind, hospitable, wine loving stereotypical Italian. Too bad you didn't start your PhD earlier, but for the two years that we were at SSB together, you were definitely one of the best party-goers and funnier people there! Don't drink too much though, you have a PhD to finish…

Linde, you have the ability to make me really, really, really uncomfortable. But even then, you are still a fun and nice person to hang out with! You're a very hard-working and pro-active woman so I'm sure you will have success in your career. You actually have a life, unlike most people in SSB, so I can count the times you had a drink

with us on my hands. But those were good times regardless! Keep on doing what you do, darlie.

Nikolas! We had really good times together, especially outside of work playing HotS! I haven't been that active with games lately, but I'm sure we will pick it up sometime in the future. I pity the girls at your new job. They have to work together with such a charmer, knowing that they can't all have you… But that is life. Good luck with your defense coming up soon, and if you're ever in Switzerland let me know!

Javi, bastardo. You already left a long time ago and we didn't really work together, but we always had good times together outside of work! Although, I have to confess, the first year or so I had a really hard time understanding you… I hope you and **Cata** are enjoying Luxembourg and wherever you are going next!

Maria, you were always there when I needed help or advise in science! It is a pity we never worked together, because I think I could have learned a lot from you, as can be seen from everyone in SSB that did work together with you! I am glad you got a tenure track position, and wish you all the luck in your career.

Edoardo, like Maria you were always there when I needed (statistical) advise, and it is a pity we never worked together. I also wish you all the best in your career.

Niels, you are definitely a weird guy too. And you are not even trying! Vegetarian, yoga, meditation… I just don't get it man! I do get that you are a genuinely nice guy, though! Because you are more settled than most of us, we never hang out *too* much outside of work. We still got invited to your place every now and then, and of course I can't forget your wedding. I wish you, your wife, and your kids all the best in life, wherever that may take you.

Stamatios, I tried play fighting you a few times during the start of our PhD's… I learned to not do that. Thanks for being an excellent iGEM supervisor with us, I had a great time. All the best in your scientific career, the ambition is definitely there!

Rik, aka Jesus. You are more of a quiet type in a group, but we had some fun coffee breaks together regardless! A very funny, but at the same time hard-working guy, I'm sure you will have a successful career.

Anna, I can basically just type: "same as Rick", because it is true. More quiet in bigger groups, but very fun in smaller ones. Once Nikolas and I will pick up our gaming rhythm again, you have to join us. We won't give you a choice.

Melanie, you were already a presence when you were a student, I am sure you are even more now! I trust on you to lead all new SSB employees to party greatness…

Nong, the selfie queen! Always in for lunch with us. A bit hard to talk over some of the loud people in SSB, but always a nice person!

Rita, sadly we never worked together. Mostly, of course, because I am a computational biologist and you're in the wet lab. All the best during the rest of your career!

Marta, like with Rita and Nong, you were in the wet lab and I wasn't. But, we supervised iGEM together! Thanks for that experience, I will remember it for a long time!

Tom, like with the others above, wet lab, dry lab… But the few times I needed you, you helped me out. So thank you for that!

Carolien, the secretary for most of my PhD. Everything went very smoothly while you were there. I don't think your influence can be overestimated!

Kal, always a myth to me. I never know what you're thinking, doing, or what you're going to do. I have no clue where you are right now, and I have no clue where you will be in the future. I never had a clue if your stories during coffee breaks were going to be funny, or if I would roll my eyes. But you were always present when drinks were to be had, and you were definitely one of the fun guys! Good luck at… wherever you are.

Apart from the people at SSB, I want to thank many others. First, I have been party of two amazing projects. The **PhD trip** to California was one of the best experiences in my life so far, and organizing it has given me a lot of valuable experience (and a better c.v.)! I want to thank **Alex K., Ana, Yue, Ruben, Jasper S., Kees, and Nico** for having a flawless PhD trip and a flawless organization. Did I mention it was flawless? I am not even exaggerating. Thanks guys!

I already mentioned **iGEM** a few times, but I will do it here specifically. I want to thank all the students and supervisors for the amazing 2016 iGEM team! We had ups and downs, like every iGEM team, but the result was one we could not have dreamed of! Furthermore, I feel privileged to have been allowed to go to Boston with you guys. It has been one of the best experiences in my life! Thank you, **Thomas, Bel, Carina, Lisa, Mark, Ronald, Angelina, Marijn, Jaccoline, Remco, Mario, Linea, Emma, Rob, Alex K., Franklin, Kees, Marta, and Stamatios**.

On the topic of students, I want to thank my three bachelor and masters students during my PhD: **Floris, Tohm, and Ronald**. My development as a scientists would not have been the same without having you guys to help me along the way.

And last, but definitely not least, I want to thank the **'SHOOOOTS'** group and everyone else I haven't mentioned before that I spent time having drinks with. I wrote a list of names, but considering I saw

most of you only when absolutely drunk, I will probably forget a lot of people. Sorry for that!

Irene, I wrote you down first on my list. And there is probably a reason for that. You are the most annoying person I met in Wageningen. But also one of the most fun. I always had great times drinking with you, and I know you did with me too, don't deny it. See you soon!

Peer, second on my list granted mostly because I thought of Irene first. But as one of the few other Dutch guys in the crew, and even one from Brabant that studied in Nijmegen, you were one of the chosen ones to actually get most of my jokes. For that I am eternally grateful.

Monika, the craziest girl from Wageningen. I heard from everyone that Wageningen got super boring after we left, and I am sure you are a big reason for that. I lost all my hair because of alcohol poisoning due to all the SHOOOOTS you fed us. Thanks for that…

Jeroen, you appear quiet, but are extrovert at the same time. Calm, but fun at the same time. We didn't share as many drinks with each other as some others on this list, but more than enough to have some great memories!

Rebecca, of course you are high on my list. Wageningen lost a gem after you left for Delft. Humour is often a language thing, and because of that it was easy to laugh with you in English. It is a pity I am so horrible at keeping connections with people the moment they move more than 5km away from me, but I hope we will share a drink sometime in the future!

CataLINA. Mamita rica! It took me a long time to start saying Cata instead of Catalina even after you insisting on being called Cata every single time. What can I say, another crazy girl which I drank way too many shots with. You and Javi were the first to visit me in

Lausanne, I really appreciate that. I am sure we will see each other often enough in the future!

Lara, sometimes we had clashing personalities, but we definitely had good times together! I hope you are having a great time in Ireland. A great place to drink alcohol. Maybe not shoooots, but beer and whiskey surely.

Jueeli, not super crazy, never super drunk, but definitely never quiet. Always fun to hang out with you whenever you joined us for drinks!

Sven, what can I say. First of all, we have so many great memories thanks to you always having a camera with you. Some things which I'd rather don't remember! Probably the craziest drinker of us all. It's a good thing you got a job that sometimes accommodates that! See you soon!

Cristina. Always in for social events! I hope you are not rotting away in the boring northern part of my country…

Alex U., hit and miss. Sometimes you can drink with us and be O.K. And sometimes you drink one beer and you're completely wasted! We're always in for a surprise when you're with us, that's for sure.

Juanan, no clue why you are so low on my list, you should a 100% be somewhere at the top! So many good memories with you, papito rico! You are one of the reasons why I have such a good image of the Spanish, even with people like Irene! I am proud of my Spanish heritage!

Yuan, did I ever see you drunk? Settled down once you got a cat. You're a real mom now, so you are forgiven. Still very fun to hang around with whenever we get the chance! See you soon. Or I guess, actually on the way to my defense. Since I'm driving with you. Let's not get drunk in the car…

To anyone that I missed: sorry! In true style, I wrote this at the last possible moment. I will apologize to you in person if you come to my PhD defense party. You can even have some drinks for free ;).

Overview of completed training activities

Discipline specific	
Linear and integer programming	2013
Training workshop interdisciplinary life sciences	2013
SB@NL 2014	2014
NBIC 2014	2014
SPLASH 4th general assembly	2015
Data integration in the life sciences	2015
BioSB 2016	2016
ECCB 2016	2016
SPLASH 5th general assembly	2016
iGEM 2016 world jamboree supervisor	2016
BioSB	2017
General courses	
VLAG PhD week	2013
Scientific writing	2014
Presentation skills	2014
Teaching and supervising thesis students	2015
Writing grant proposals	2017
Optionals	
Preparation of research proposal	
Weekly group meetings	
Seminar series	
PhD trip 2015	2015
PhD trip 2015 organization	2015

List of publications

Reijnders, M. J., van Heck, R. G., Lam, C. M., Scaife, M. A., dos Santos, V. A. M., Smith, A. G., & Schaap, P. J. (2014). Green genes: bioinformatics and systems-biology innovations drive algal biotechnology. *Trends in biotechnology*, *32*(12), 617-626.

Reijnders, M. J., Carreres, B. M., & Schaap, P. J. (2015). Algal omics: The functional annotation challenge. *Current Biotechnology*, *4*(4), 457-463.

Schulze, S., Urzica, E., Reijnders, M. J., Geest, H., Warris, S., Bakker, L. V., ... & Hippler, M. (2017). Identification of methylated GnTI-dependent N-glycans in Botryococcus brauni. *New Phytologist*, *215*(4), 1361-1369.

The research described in this thesis was financially supported by the IPOP program Systems Biology of Wageningen University & Research, and by the European Community's Seventh Program for research, technological development and demonstration under grant agreement No Fp7-311956

Cover design by Matheus Guimarães
Printed by Proefschriftmaken on FSC-certified paper