Green genes: bioinformatics and systems-biology innovations drive algal biotechnology

Maarten J.M.F. Reijnders1, Ruben G.A. van Heck1, Carolyn M.C. Lam1,2, Mark A. Scaife3, Vitor A.P. Martins dos Santos1,2, Alison G. Smith3, and Peter J. Schaap1

1Laboratory of Systems and Synthetic Biology, Wageningen University, Dreijenplein 10, Building number 316, 6703 HB Wageningen, The Netherlands
2LifeGlimmer GmbH, Markelstrasse 38, 12163 Berlin, Germany
3Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK

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.

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 [1,2]. 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 [3,4], owing to rapid growth rates, amenability to large-scale fermentation, and the potential for sustainable process development [5].

Algae as a source of biofuel molecules, such as triacylglycerides (TAGs), the precursor for biodiesel [6], have been a focus in recent years, with potential yields an order of magnitude greater than competing agricultural processes [7]. Evaluations of current technologies demonstrate that microalgae are commercially feasible for biofuel production, but are not yet cost-competitive with petroleum products [8,9], 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 [8]. 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 bio-synthesis of TAG and cell growth is therefore a severely limiting factor [10]. 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 [11]. These carbohydrates can be degraded to fermentable sugars for bioethanol production [12], 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 [3], 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 [13]. 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.

Annotation challenges for microalgae

The nuclear genome of the green alga 
*Chlamydomonas reinhardtii*, sequenced in 2007 [1], 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 [1], 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 [14], 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) [15], which applies the one-to-many recognition strategy by calculating the statistical significance of all candidate homologous genes found by BLAST [16] and HMMER [17], 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 [18], transcript expression patterns [18], phylogenetic trees [19], compartmentalization information [20], and literature [21] are also taken into account. FFPred2 from UCL–Jones [20] is the prime example of such a homology-independent functional annotation algorithm.

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Success rate*</th>
<th>Computational speed</th>
<th>Availability</th>
<th>Additional notes</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard BLAST</td>
<td>Limited</td>
<td>Fast</td>
<td>Online/offline</td>
<td>Dependent on global sequence similarity for success</td>
<td>[16]</td>
</tr>
<tr>
<td>HMMER</td>
<td>Moderate</td>
<td>Fast</td>
<td>Online/offline</td>
<td>Family-wise alignment method</td>
<td>[17]</td>
</tr>
<tr>
<td>InterProScan</td>
<td>Moderate</td>
<td>Slow</td>
<td>Online/offline</td>
<td>Family-wise alignment method</td>
<td>[72]</td>
</tr>
<tr>
<td>FFPred2</td>
<td>High</td>
<td>Slow</td>
<td>Limited online/offline</td>
<td>Algorithms currently trained on non-algal datasets</td>
<td>[20,23]</td>
</tr>
<tr>
<td>Argot2</td>
<td>High</td>
<td>Moderate</td>
<td>Limited online</td>
<td>Initial selection is dependent on BLAST and HMMER output</td>
<td>[15]</td>
</tr>
</tbody>
</table>

*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) [15]. 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.
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 [22]. 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 [23,24], but whilst studies have shown that this is a feasible option [25], 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] [15], TargetP [26], SignalP [27], PSORTb [28], and PredAlgo [29]. 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 [30], or the collection of roughly 1000 chloroplast-localized proteins from C. reinhardtii [31], the algorithm is trained with a relatively small C. reinhardtii dataset [29]. 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 [32], a C. reinhardtii-specific pathway/gene database of the MetaCyc [33] facility for metabolic pathway analysis. A peptide database, ProMEX, is available that contains over 2000 C. reinhardtii peptides which are usable for proteomics analysis [34]. In addition, the Augustus tool, which is commonly used for prediction of eukaryotic genes [35], has a tailor-made section for C. reinhardtii. Finally, the Algal Functional Annotation Tool [36] 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) [37], 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.

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 analyzed. 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 behavior. These methods include the widely employed flux balance analysis (FBA) [38] and flux variability analysis (FVA) [39], but also methods integrating fluxomic, transcriptomic, or proteomic data (Box 2) [38]. For an extensive overview of mathematical methods using metabolic models we refer to Zomorrodi et al. [40]. 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, incon-
sistencies 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* [41,42] (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 [41]. 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 [42]. These updated reactions and annotations [41,42] were subsequently stored in ChlamyCyc [32], 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 [43]. This model was greatly expanded in iRC1080 in 2011 and additional ORFs were validated [44]. 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 [44].

Recently GSMMs for *Ostreococcus tauri* and *Ostreococcus lucimarinus* have been constructed [45] (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 [45]. The gap-filling method designed in [42] was subsequently employed to find suitable reactions for the production of these metabolites [45].

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 [46,47], and to contrast mutant strains [44]. *C. reinhardtii* metabolic models were also used to determine how the quantity of light [44,48,49] and its spectral composition [44] affect metabolism. Of particular interest is the possibility to predict an optimal light spectrum for a given metabolic goal [44]. 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 [50]. Potentially, these algae and associated bacteria can be modeled simultaneously to deduce their relationship, as has been done for other microbial communities [51,52].

The most comprehensive algal metabolic models to date are iRC1080 [44] 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

### Table 2. A list of selected industrially useful microalgae

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome size (Mb)</th>
<th>Available proteins</th>
<th>Reported industrially relevant characteristics</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Chlamydomonas reinhardtii</em></td>
<td>120</td>
<td>15 144</td>
<td>Model system for unicellular green algae</td>
<td>[73]</td>
</tr>
<tr>
<td><em>Monoraphidium neglectum</em></td>
<td>68</td>
<td>16 761</td>
<td>Up to 21% dry weight neutral lipid under nitrogen starvation</td>
<td>[4]</td>
</tr>
<tr>
<td><em>Nannochloropsis gaditana</em></td>
<td>34</td>
<td>15 361</td>
<td>Can produce high amounts of ω-3 long-chain polyunsaturated fatty acids</td>
<td>[74]</td>
</tr>
<tr>
<td><em>Nannochloropsis oceanica</em></td>
<td>28</td>
<td>242</td>
<td>Up to 50% dry weight oil content</td>
<td>[42]</td>
</tr>
<tr>
<td><em>Pheoeactum tricornutum</em></td>
<td>27</td>
<td>10 673</td>
<td>Can produce antibacterial fatty acids (9Z)-hexadecenoic acid (palmitoleic; C16:1 n-7) and (6Z, 9Z, 12Z)-hexadecatrienoic acid (HTA; C16:3 n-4)</td>
<td>[50]</td>
</tr>
<tr>
<td><em>Chlorella variabilis</em></td>
<td>46</td>
<td>9831</td>
<td>The first sequenced <em>Chlorella</em> genome</td>
<td>[75]</td>
</tr>
<tr>
<td><em>Ostreococcus tauri</em></td>
<td>12.6</td>
<td>9050</td>
<td>Smallest sequenced microalgal genome with simple cellular structure</td>
<td>[48]</td>
</tr>
<tr>
<td><em>Chlorella protothecoides</em></td>
<td>22.9</td>
<td>7039</td>
<td>Up to 55% dry weight lipid content in heterotrophic growth</td>
<td>[46,77]</td>
</tr>
<tr>
<td><em>Chlorella vulgaris</em></td>
<td>N.a.</td>
<td>292</td>
<td>Up to 42% lipid content in photobioreactor with artificial waste water</td>
<td>[78]</td>
</tr>
<tr>
<td><em>Dunaliella salina</em></td>
<td>N.a.</td>
<td>238</td>
<td>Up to 10% carotenoids in dry weight; 90% β-carotene in carotenoids</td>
<td>[79]</td>
</tr>
<tr>
<td><em>Haematococcus pluvialis</em></td>
<td>N.a.</td>
<td>60</td>
<td>Highest reported yield of antioxidant astaxanthin (3.8% dry weight)</td>
<td>[80]</td>
</tr>
<tr>
<td><em>Botryococcus braunii</em></td>
<td>~166–211</td>
<td>30</td>
<td>Up to 57% total lipids in dry weight</td>
<td>[81–83]</td>
</tr>
<tr>
<td><em>Neochloris oleoabundans</em></td>
<td>N.a.</td>
<td>0</td>
<td>Up to 56% total fatty acids in dry weight under nitrogen-deprivation</td>
<td>[10]</td>
</tr>
</tbody>
</table>

*Genome size or characteristics are according to NCBI unless otherwise specified.

*Estimated protein numbers are according to UniProt unless otherwise specified.

*N.a., not available.
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 [55,56], 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 [56], 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 [57]. Another study has modeled different pathways for the process of carbon fixation [58] as a means to overcome the low oxygenase activity of Rubisco [59]. Bar-Even et al. [58] 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 [10]. 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
<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Species</th>
<th>Parent model (if any)</th>
<th>Year</th>
<th>Species</th>
<th>Parent model (if any)</th>
<th>Year</th>
<th>Species</th>
<th>Parent model (if any)</th>
<th>Year</th>
<th>Species</th>
<th>Parent model (if any)</th>
<th>Year</th>
<th>Species</th>
<th>Parent model (if any)</th>
<th>Year</th>
<th>Species</th>
<th>Parent model (if any)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christian et al. [42]</td>
<td>2009</td>
<td>C. reinhardtii</td>
<td>/C0/C0/C0/C0</td>
<td>2011</td>
<td>C. reinhardtii</td>
<td>Boyle and Morgan</td>
<td>2011</td>
<td>C. reinhardtii</td>
<td>iAM303</td>
<td>2011</td>
<td>C. reinhardtii</td>
<td>Boyle and Morgan</td>
<td>2011</td>
<td>C. reinhardtii</td>
<td>Kliphuis et al. [48]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Overview of metabolic models of microalgae

- 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.
- The category ‘others’ refers to the following compartments: flagellum, Golgi apparatus, thylakoid lumen, nucleus, and eyespot.
- N.a., not available.
- Gene information not available from model files nor supplementary documents.
of nutrient limitation [60]. 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 [61]. 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.
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.

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 Programme (FP7) project SPLASH (Sustainable Polymers from Algae Sugars and Hydrocarbons), grant agreement number 311956.

References
24 Rodgers-Melnick, E. et al. (2013) Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS. BMC Genomics 14, 608
46 Gomes de Oliveira Dal'Molin, C. et al. (2011) AlgaGEM – a genome-scale metabolic reconstruction of algae based on the Chlamydomonas reinhardtii genome. BMC Genomics 12 (Suppl. 4), S5
50 Rivas, M.O. et al. (2010) Interactions of Botryococcus braunii cultures with bacterial biofilms. Microb. Ecol. 60, 628–635
60 Guarnieri, M.T. et al. (2013) Proteomic analysis of Chlorella vulgaris: potential targets for enhanced lipid accumulation. J. Proteomics 93, 245–253
70 Rasala, B.A. et al. (2012) Robust expression and secretion of xylanase 1 in Chlamydomonas reinhardtii by fusion to a selection gene and processing with the FMDV 2A peptide. PLoS ONE 7, e43549
73 Bogen, C. et al. (2013) Reconstruction of the lipid metabolism for the microalgae Monoraphidium neglectum from its genome sequence reveals characteristics suitable for biofuel production. BMC Genomics 14, 926
76 Gao, C. et al. (2014) Oil accumulation mechanisms of the oleaginous microalga Chlorella protothecoides revealed through its genome, transcriptomes, and proteomes. BMC Genomics 15, 582
87 Yizhak, K. et al. (2010) Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Bioinformatics 26, 1255–1260