Green genes: bioinformatics and systems biology innovations drive algal biotechnology

Authors' names:

Maarten J.M.F. Reijnders[1], Ruben 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], Peter J. Schaap[1]

Authors' addresses:

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

[2]LifeGlimmer GmbH, Markelstr. 38, D-12163 Berlin, Germany.

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

Corresponding author: Schaap, Peter J. (peter.schaap@wur.nl)

Abstract (120 words)

Many species of microalgae possess the capacity to 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 improve productivity, but the exact metabolic pathways and identity 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 multi-disciplinary integration of methods will provide synergy for a systems-level understanding of microalgae, and thereby speed up the improvement of industrially valuable strains. In this review we highlight recent advances, challenges, their application in microalgae research, and potentials.

Introduction

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, which is reflected in the genetic diversity that exists both between species, and within a single genome, where a seeming mosaic of genetic origins can be observed [1]. Mining the genomes of these organisms is 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 [2, 3], due to rapid growth rates, amenability for large scale fermentation, and the potential for sustainable process development [4].

Algae as a source of biofuel molecules, such as triacylglycerides (TAGs), the precursor for biodiesel [5], have been a focus in recent years, with potential yields an order of magnitude greater than competing agricultural processes [6]. Currently, in order 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. This trade-off between biosynthesis of TAG and cell growth is therefore a severely limiting factor [7]. 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 on a systems level. For example, polysaccharides such as starch and cell wall materials can be used for biotechnological applications [8]. These carbohydrates can be degraded to fermentable sugars for bio-ethanol production [9], 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 [2], 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 performed. Yet our understanding of metabolic pathways within these microalgae remains limited [10]. A significant knowledge gap needs to be filled between omics data, the annotation thereof and our systems level understanding. This will allow the conversion of these resources into useable genome scale models, and provide the basis for effective metabolic engineering, synthetic biology and biotechnology. Here we consider novel approaches to improve annotation of algal omics data, the establishment of

genome-scale metabolic models, and ways to integrate these for effective exploitation of microalgae for biotechnology.

Annotation challenges for microalgae

The nuclear genome of the model green alga *Chlamydomonas reinhardtii*, sequenced in 2007 [1], is approximately 120 Mb and encodes ~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 experimentally validated functions 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 other organisms, including *A. thaliana,* other plants and microbes [1], using BLAST or family-wise alignment methods such as HMMER and InterProScan (Table 2). 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 annotation of distantly related *C. reinhardtii* genes, but becomes even more problematic when an *in silico* derived annotation of *C. reinhardtii* is subsequently used for the annotation of other algal species. This is because two algal species can be more diverse than any two plant species, for example. Therefore, these methods, which are highly suitable for high-throughput analysis due to their simplicity, are not suitable for accurate in-depth annotation of algal genomes. In the Critical Assessment of protein Function Annotation (CAFA) experiment [11], the accuracy of more advanced function annotation algorithms was assessed. The CAFA concluded that 33 of 54 tested function annotation algorithms outperformed the standard BLAST-based method (Table 2). The substantial improvement can be explained by the fact that these 2nd generation methods do not apply the one-to-one

recognition principal, but rather use a one-to-many recognition strategy to increase their success rate and include context aware principles for annotation. An example is Argot2 (Box 1) [12], which applies the one-to-many recognition strategy by calculating the statistical significance of all candidate homologous genes found by BLAST [13] and HMMER [14], combined with an assessment of semantic similarities of associated GO terms. In a context aware multi-level approach, annotation is not merely based on sequence similarity, but also other factors such as protein-protein interactions [15], transcript expression patterns [15], phylogenetic trees [16], compartmentalization [17], and literature [18] are taken into account. FFPred2 from UCL-Jones [17] is the prime example of such a homology-independent function annotation algorithm.

Advanced multi-level annotation methods effectively increase the recall of function prediction while maintaining a reasonable precision. The challenge in genome annotation of microalgae lies in the small number of experimentally validated algal genes and the lack of algae-specific contextual data such as protein interaction data. This results in a relatively low number of genes 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, whilst a consensus prediction among the different methods improves the accuracy of the annotation [19]. Due to their simplicity and speed, 1st-generation methods can be used for initial high-throughput analysis of a large set of genes. 2nd-generation methods can then be used for a refined analysis of these genes. However, in order to utilize these advanced methods fully, a significant amount of experimentally determined contextual data is required. Whilst increasing amounts of gene expression data are being generated, there is still little structural and protein interaction data available for algae. In the absence of such experimental facts, it is still possible to generate this contextual information by *in silico* prediction methods

[20, 21]. Although studies have shown that this is a feasible option [22], caution is necessary, since 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) [12], TargetP [23], SignalP [24], PSORTb [25] and PredAlgo [26]. The latter of which 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, due to the limited number of algal proteins with a known cellular localization, the algorithm is trained with a relatively small *C. reinhardtii* data set. This raises questions regarding reliability for other algal species that are more distantly related. 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 give the means to share data easily so 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. Additionally, tools and databases that do more than storing the available sequencing data are needed. There are a small number of tools available, although these are often limited to *C. reinhardtii.* One such tool is ChlamyCyc [27] a *C. reinhardtii* specific pathway/genome database of the MetaCyc [28] facility for metabolic pathway analysis. Additionally, the Augustus tool, which is commonly used for prediction of eukaryotic genes [29], has a tailor-made section for *C. reinhardtii.* Finally, the Algal Functional Annotation Tool [30] 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) [31], 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 summarizing 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 1). Such a web-portal would provide the means for new and existing tools specifically useful for algal species to facilitate exposure to a broad audience. Additionally, it could act as a hosting platform for small but useful tools like 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 in reducing the redundancy that is often seen in biology and bioinformatics.

Table 1: A list of selected industrially useful microalgae.

Species	Genome size‡ (Mb)	Proteins in UniProt‡	Characteristics‡	Ref‡
Chlamydomonas reinhardtii	120	15,144	• Model system for unicellular green algae • Rapid growth	
Monoraphidium neglectum	68	16,761	• Biofuel production candidate	[32]
Nannochloropsis sp.	44	16,226	• Produces high amounts of omega-3-long-chain polyunsaturated fatty acids	
Phaeodactylum tricornutum	27	10,673	• Production of antibacterial fatty acids	[33]
Chlorella variabilis	46	9,831	• Contains several essential nutrients • Rich source of lutein	
Ostreococcus tauri	12.6	9,050	• Smallest microalgal genome	
Chlorella vulgaris	n.a.	292	• High lipid content under nitrogen limitation	[34]
Dunaliella salina	n.a.	238	• High concentration of beta-carotene	[35]
Chlorella protothecoides	n.a.	96	• High lipid content in heterotrophic growth [3] • Highest published biomass yield [36]	[3], [36]
Haematococcus pluvialis	n.a.	60	• Antioxidant astaxanthin production	[37]
Botryococcus braunii	~166 – 211 [38]	30	• High levels of liquid hydrocarbons and exopolysaccharides [39]	[38], [39]
Neochloris oleoabundans	n.a.	0	• High lipid content	[7]

‡: Genome size, estimated protein numbers and characteristics are according to NCBI and UniProt, unless otherwise specified.

n.a.: Not available.

Table 2: Comparison of the features of commonly used functional annotation tools.

Methods	Success rate*	Computational speed	Availability	Additional notes	Ref
Standard BLAST	Limited	Fast	Online/ offline	• Dependent on global sequence similarity for success • Suitable for high throughput analysis	[13]
HMMER	Moderate	Fast	Online/ offline	• Family-wise alignment method • Suitable for high throughput analysis	[14]
InterProScan	Moderate	Slow	Online/ offline	• Family-wise alignment method • Uses pre-computed protein domains	[40]
FFPred2	High	Slow	Limited online/ offline	• Algorithms currently trained on non-algal datasets • Not suitable for high throughput analysis	[17, 20]
Argot2	High	Moderate	Limited online	• Initial selection is dependent on BLAST and HMMER output • Additionally predicts compartmentalization • User-friendly interface	[12]

* For distantly related sequences

Box 1: Argot2

One of the top performers in the CAFA experiment is Argot2 [12]. It stands out when it comes to simplicity, as well as its incorporation of BLAST and HMMER. This method combines an easy interface with multi-layer analysis, making it a perfect starting point for biologists who want 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 GO terms for each query sequence. The user can choose to perform this entire process online at the Argot2 webserver, limited to a hundred sequences per query. Alternatively, if the BLAST and HMMER steps are performed by the user locally and provided to the webserver, over a thousand 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.

Box 2: Flux analysis in microalgae

Flux balance analysis (FBA) [41] is the most commonly applied method to simulate metabolism in genome-scale metabolic models. It identifies a theoretically optimal use of metabolic capabilities for a selected metabolic objective, in a specific environment. As 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 *e.g.* how a microalga can achieve maximal growth under defined illumination. In addition, disabling metabolic capabilities associated with a gene allows simulation of mutant strains. FBA can thus assess the potential of different strains and different environmental conditions. In order 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. (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 of, for example, algal biomass production. Together, the *S* matrix, the constraints, and the objective function form a linear programming problem:

$$\max(\boldsymbol{x} * \boldsymbol{c})$$

$$s.t. \quad S * \boldsymbol{x} = \boldsymbol{0}$$

$$\boldsymbol{x} \geq \boldsymbol{lb}$$

$$\boldsymbol{x} \leq \boldsymbol{ub}$$

With: $\boldsymbol{x}$ — flux vector, $\boldsymbol{c}$ —objective vector, $\boldsymbol{0}$ — a null vector ensuring steady state, and $\boldsymbol{lb}/\boldsymbol{ub}$ — lower/upper bounds for each reaction. The vector $\boldsymbol{x}$ represents a flux distribution with the theoretically maximal value for the metabolic objective. However, due to the presence of alternative/cyclic pathways, there are often alternative flux distributions with equally high values for the objective function. Flux variability analysis [34] explores for each reaction to what extent the flux can vary while permitting only a small reduction in the obtained objective value. In addition, experimental data can be used to provide additional constraints. For example, ^{13}C-labelling experiments provides experimentally measured fluxes as inputs for the model simulations [42, 43]. Several FBA-based methods also facilitate the integration of transcriptomic and proteomic data with metabolic models to constrain reactions based on the gene expression levels [44]. Thereby, flux distributions are identified which are most consistent with the expression data [45]. Metabolic models are thus integrated with, and their predictions compared to, experimental data yielding new insights in metabolic functioning.

Understanding algal metabolism on 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 genome 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 and the associated genes and constraints, e.g. reaction directionalities and rate limits. Due to their comprehensive representation of metabolism, metabolic models form the basis for a large and diverse set of mathematical methods predicting metabolic behavior. These methods include the widely employed Flux Balance Analysis (FBA) [41] and Flux Variability Analysis (FVA) [46], but also methods integrating fluxomic, transcriptomic, or proteomic data (see Box 2) [41]. For an extensive overview of mathematical methods using metabolic models we refer to Zomorrodi *et al.* [47]. Here, we focus 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* [48, 49] (Figure 2). Metabolites found experimentally but not in the models initiated pathway elucidation and identification of the corresponding genes, and thereby led to an improved genome annotation [48]. 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 [49]. These updated reactions and annotations [48, 49] were subsequently stored in ChlamyCyc [27], 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 genome annotation, and to additional support for the reactions included in the model [50]. This model was greatly expanded in iRC1080 in 2011 and additional ORFs were validated [51]. 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 varying experimental conditions, although these predictions remain to be verified [51]. Recently GSMMs for *Ostreococcus tauri* and *Ostreococcus lucimarinus* have been constructed [52] (Figure 2), 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 [52]. The gap-filling method designed in [49] was subsequently employed to find suitable reactions for the production of these metabolites [52].

It is well recognized that the exact choice of growth conditions is highly important to attain desired metabolic activities. Metabolic models can explore how different growth conditions affect metabolism and 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 [53], to investigate how metabolite production is influenced [53, 54], and to contrast mutant strains [51]. *C. reinhardtii* metabolic models were also used to determine how quantity of light [51, 55, 56] and its spectral composition [51] affect metabolism. Of particular interest is the possibility to predict an optimal light spectrum for a given metabolic goal [51]. 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 [57]. Potentially, these algae and associated bacteria can be modeled simultaneously in order to deduce their relationship, as has been done for microbial communities [58, 59].

The most comprehensive algal metabolic models to date are iRC1080 [51] and AlgaGEM [53], which are GSMMs and account for various cellular compartments. However, they vary in degree of compartmentalization (Figure 2). In iRC1080, half (865/1730) of the non-transport reactions occur in cellular compartments other than the cytosol. In contrast, this is only about 12% (201/1617) for AlgaGEM. This reflects that independently generated GSMMs for the same organism can differ significantly in their representation of metabolism, as 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 genome data for the alga in question, but employing the strategies and tools developed for *C. reinhardtii*, as has been done for *Ostreococcus* [52]. Ultimately, GSMMs of various microalgae will be valuable for designing strategies that increase the production of compounds of interest [47, 60]. 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 [61], will pave the way for model-driven engineering of algal species.

Figure 1: Overview of metabolic models of microalgae. Green boxes represent *C. reinhardtii* GSMMs, the red box represents an *Arabidopsis thaliana* GSMM associated with a *Chlamydomonas reinhardtii* GSMM, and the blue box represents two *Ostreococcus* GSMMs. A connection between two GSMMs indicates that the former was used in the reconstruction of the latter. The boxes are annotated with the model names if available and otherwise with the author name(s): Christian *et al.* [49], Boyle & Morgan [54], AraGEM [62], iAM303 [50], Cogne *et al.* [56], Kliphuis *et al.* [55], AlgaGEM [53], iRC1080 [51], Krumholz *et al.* [52]. The numbers in each box indicate the total number of reactions (R), total number of genes (G), unique decompartmentalized metabolites (M), and biological cellular compartments (C) found in available model files. The pie charts depict the distribution of biochemical reactions among different compartments as well as compartment-spanning transport reactions (reaction categories are shown in the legend). The category 'others' refers to the following compartments: flagellum, Golgi apparatus, thylakoid lumen, nucleus, and eyespot.

*: Additional information obtained from authors.

**: Gene information not available from model files.

Integrating bioinformatics and modeling for algal biotechnology

Improvements in algal annotations will need to interact closely with systems modeling of the metabolic and regulatory networks in order to refine our understanding of the capabilities of a specific alga, and to provide a basis for applications in biotechnology. Figure 2 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 identification of genes and proteins (step 2). After 1st-generation high-throughput functional annotation (step 3), a refinement step using 2nd-generation function 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 refinement of annotation is required. Due 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, which 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.

The GSMMs provide a basis for both computational and lab-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 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 [63, 64], 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 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, and increasing the surface area to volume ratio required for maximum productivity. Truncated light-harvesting chlorophyll antenna size (*tla*) mutation with reduced antenna size in *C. reinhardtii* has been shown to improve solar energy conversion efficiency and photosynthetic productivity in mass culture and bright light [65]. Another study has modeled different pathways for the process of carbon fixation [66], as a means to overcome the low oxygenase activity of Rubisco [67]. Bar-Even *et al.* [66] computationally identified alternative carbon fixation pathways by using approximately 5,000 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 three times more efficient than the conventional Calvin-Benson cycle. Using an algal GSM to study these pathways would help to understand 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 [7]. This also triggers a reduction in photosynthetic membrane lipids and cessation of cell growth. The link between TAG accumulation and macronutrient stress has been investigated using a systems approach, leading to the identification of a putative N-triggered transcription factor in *C. reinhardtii*, the overexpression of which resulted in about 50% increase in lipid production under specific experimental conditions [68]. In another approach, in the diatom, *Thalassiosira pseudonana*, TAG production was increased via an RNAi knockdown

strategy targeting not the biosynthesis of lipids, or the production of competing energy sinks, but instead targeting lipases, involved in glycerolipid catabolism [69]. 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 two fold over chemically inducible and constitutive gene expression [70]. 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.

In contrast to the use of bacteria and yeasts for industrial production, algal biotechnology is in its infancy. Alongside genome sequence information, a key requirement is the ability to carry out genetic transformation, and while this is only routine for *C. reinhardtii,* and the diatoms *P. tricornutum* and *Thalassiosira pseudonana,* in the last few years there has been a rapid increase in published methods for transformation of several species of industrial interest including *Nannochloropsis* sp. [71]. Moreover, the ability to engineer the chloroplast genome offers considerable opportunities for metabolic engineering, given the focus of this organelle on biosynthesis [72]. But 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 a number of well-established systems for inducible gene expression in *C. reinhardtii,* most notably promoters that are regulated in response to nitrate (*NIT1* or *NIA1*) [73] or , copper (*CYC6*) [74]. More recently, vitamin responsive *cis*-elements have been identified, namely a cobalamin (vitamin B_{12}) responsive promoter [75], as well as a thiamine (vitamin B_1)

responsive riboswitch [76], have been demonstrated as useful regulatory tools. Vitamins present advantages of being benign, cheap and effective at low concentrations. However, the majority of these have been discovered by coincidence rather than 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. [77]. Further facilitation of transgene expression comes from the use of 2A peptides [78], 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* [72, 79]. In spite of these developments, progress remains parallel in nature, and heavily focused upon the development of *C. reinhardtii*.

For microalgae to develop as a biotechnology platform, rational design to address their current shortcomings must be achieved through the development of fit-for-purpose metabolic engineering or synthetic biology resources. 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 2) offers new perspectives for future improvements in algal biotechnology. The current prominent challenges in algal bioinformatics and genome-scale modeling are the foundation for overcoming the knowledge barrier to enable predictive modifications of various algal genomes in the future.

Figure 2: A multidisciplinary workflow for integrative and systematic understanding of algae.

Black arrow: *in silico* data or predictions; white arrow: experimental (wet-lab) data.

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 co-existing bacteria, is a crucial first step. Extending bioinformatics approaches for function prediction through incorporation of new methodology, integrated and flexible databases, and 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.

Acknowledgement

We acknowledge support from the EU FP7 project SPLASH (Sustainable PoLymers from Algae Sugars and Hydrocarbons), grant agreement number 311956.

References

1 Merchant, S.S., *et al.* (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318, 245-250

2 Borowitzka, M.A. (2013) High-value products from microalgae — their development and commercialisation. *J. Appl. Phycol.* 25, 743-756

3 Scott, S.A., *et al.* (2010) Biodiesel from algae: challenges and prospects. *Curr. Opin. Biotechnol.* 21, 277-286

4 Wijffels, R.H., *et al.* (2010) Microalgae for the production of bulk chemicals and biofuels. *Biofuels, Bioprod. Biorefin.* 4, 287-295

5 Merchant, S.S., *et al.* (2012) TAG, you're it! *Chlamydomonas* as a reference organism for understanding algal triacylglycerol accumulation. *Curr. Opin. Biotechnol.* 23, 352-363

6 Mata, T.M., *et al.* (2010) Microalgae for biodiesel production and other applications: a review. *Renewable Sustainable Energy Rev.* 14, 217-232

7 Klok, A.J., *et al.* (2013) Simultaneous growth and neutral lipid accumulation in microalgae. *Bioresour. Technol.* 134, 233-243

8 Busi, M.V., *et al.* (2013) Starch metabolism in green algae. *Starch - Stärke* 66, 28-40

9 Ho, S.H., *et al.* (2013) Bioethanol production using carbohydrate-rich microalgae biomass as feedstock. *Bioresour. Technol.* 135, 191-198

10 Hildebrand, M., *et al.* (2013) Metabolic and cellular organization in evolutionarily diverse microalgae as related to biofuels production. *Curr. Opin. Chem. Biol.* 17, 506 - 514

11 Radivojac, P., *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods* 10, 221-227

12 Falda, M., *et al.* (2012) Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. *BMC Bioinf.* 13, S14

13 Altschul, S.F., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402

14 Finn, R.D., *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29-W37

15 Kourmpetis, Y.A., *et al.* (2010) Bayesian Markov Random Field analysis for protein function prediction based on network data. *PLoS One* 5, e9293

16 Engelhardt, B.E., *et al.* (2011) Genome-scale phylogenetic function annotation of large and diverse protein families. *Genome Res* 21, 1969-1980

17 Cozzetto, D., *et al.* (2013) Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC Bioinf.* 14, S1

18 Wong, A. and Shatkay, H. (2013) Protein function prediction using text-based features extracted from the biomedical literature: The CAFA Challenge. *BMC Bioinf.* 14, S14

19 Rentzsch, R. and Orengo, C.A. (2009) Protein function prediction — the power of multiplicity. *Trends Biotechnol.* 27, 210-219

20 Buchan, D.W., *et al.* (2010) Protein annotation and modelling servers at University College London. *Nucleic Acids Res.* 38, W563-568

21 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

22 Franceschini, A., *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808-815

23 Emanuelsson, O., *et al.* (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005-1016

24 Petersen, T.N., *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785-786

25 Yu, N.Y., *et al.* (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 1608-1615

26 Tardif, M., *et al.* (2012) PredAlgo: A new subcellular localization prediction tool dedicated to green algae. *Mol. Biol. Evol.* 29, 3625-3639

27 May, P., *et al.* (2009) ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii*. *BMC Genomics* 10, 209

28 Caspi, R., *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* 42, D459-D471

29 Keller, O., *et al.* (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 27, 757-763

30 Lopez, D., *et al.* (2011) Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data. *BMC Bioinf.* 12, 282

31 Goodstein, D.M., *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178-D1186

32 Bogen, C., *et al.* (2013) Reconstruction of the lipid metabolism for the microalga *Monoraphidium neglectum* from its genome sequence reveals characteristics suitable for biofuel production. *BMC Genomics* 14, 926

33 Desbois, A., *et al.* (2008) Isolation and structural characterisation of two antibacterial free fatty acids from the marine diatom, *Phaeodactylum tricornutum*. *Appl. Microbiol. Biotechnol.* 81, 755-764

34 Feng, Y., *et al.* (2011) Lipid production of *Chlorella vulgaris* cultured in artificial wastewater medium. *Bioresour. Technol.* 102, 101-105

35 Prieto, A., *et al.* (2011) Assessment of carotenoid production by *Dunaliella salina* in different culture systems and operation regimes. *J. Biotechnol.* 151, 180-185

36 Doucha, J. and Lívanský, K. (2012) Production of high-density *Chlorella* culture grown in fermenters. *J. Appl. Phycol.* 24, 35-43

37 Ambati, R.R., *et al.* (2014) Astaxanthin: sources, extraction, stability, biological activities and its commercial applications — a review. *Mar. Drugs.* 12, 128-152

38 Weiss, T.L., *et al.* (2011) Genome size and phylogenetic analysis of the A and L races of *Botryococcus braunii*. *J. Appl. Phycol.* 23, 833-839

39 Weiss, T.L., *et al.* (2012) Colony organization in the green alga *Botryococcus braunii* (race B) is specified by a complex extracellular matrix. *Eukaryot. Cell* 11, 1424-1440

40 Jones, P., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236-1240

41 Orth, J.D., *et al.* (2010) What is flux balance analysis? *Nat. Biotechnol.* 28, 245-248

42 Wiechert, W. (2001) ^{13}C metabolic flux analysis. *Metab. Eng.* 3, 195-206

43 Weitzel, M., *et al.* (2013) 13CFLUX2 — high-performance software suite for ^{13}C-metabolic flux analysis. *Bioinformatics* 29, 143-145

44 Jensen, P., *et al.* (2011) TIGER: toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC Syst. Biol.* 5, 147

45 Blazier, A.S. and Papin, J.A. (2012) Integration of expression data in genome-scale metabolic network reconstructions. *Front. Physiol.* 3, 299

46 Mahadevan, R. and Schilling, C.H. (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* 5, 264-276

47 Zomorrodi, A.R., *et al.* (2012) Mathematical optimization applications in metabolic networks. *Metab. Eng.* 14, 672-686

48 May, P., *et al.* (2008) Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of *Chlamydomonas reinhardtii*. *Genetics* 179, 157-166

49 Christian, N., *et al.* (2009) An integrative approach towards completing genome-scale metabolic networks. *Mol. Biosyst.* 5, 1889-1903

50 Manichaikul, A., *et al.* (2009) Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat. Methods* 6, 589-592

51 Chang, R.L., *et al.* (2011) Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism. *Mol. Syst. Biol.* 7, 518

52 Krumholz, E.W., *et al.* (2012) Genome-wide metabolic network reconstruction of the picoalga *Ostreococcus*. *J. Exp. Bot.* 63, 2353-2362

53 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

54 Boyle, N. and Morgan, J. (2009) Flux balance analysis of primary metabolism in *Chlamydomonas reinhardtii*. *BMC Syst. Biol.* 3, 4

55 Kliphuis, A.J., *et al.* (2012) Metabolic modeling of *Chlamydomonas reinhardtii*: energy requirements for photoautotrophic growth and maintenance. *J. Appl. Phycol.* 24, 253-266

56 Cogne, G., *et al.* (2011) A model-based method for investigating bioenergetic processes in autotrophically growing eukaryotic microalgae: application to the green algae *Chlamydomonas reinhardtii*. *Biotechnol. Progr.* 27, 631-640

57 Rivas, M.O., *et al.* (2010) Interactions of *Botryococcus braunii* cultures with bacterial biofilms. *Microb. Ecol.* 60, 628-635

58 Zomorrodi, A.R. and Maranas, C.D. (2012) OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.* 8, e1002363

59 Zomorrodi, A.R., *et al.* (2014) d-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synth. Biol.* 3, 247-257

60 Tomar, N. and De, R.K. (2013) Comparing methods for metabolic network analysis and an application to metabolic engineering. *Gene* 521, 1-14

61 Cho, A., *et al.* (2010) Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst. Biol.* 4, 35

62 Gomes de Oliveira Dal'Molin, C., *et al.* (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiol.* 152, 579-589

63 Chisti, Y. (2013) Constraints to commercialization of algal fuels. *J. Biotechnol.* 167, 201-214

64 Lehr, F. and Posten, C. (2009) Closed photo-bioreactors as tools for biofuel production. *Curr. Opin. Biotechnol.* 20, 280-285

65 Kirst, H.*, et al.* (2012) Truncated photosystem chlorophyll antenna size in the green microalga *Chlamydomonas reinhardtii* upon deletion of the TLA3-CpSRP43 gene. *Plant Physiol.* 160, 2251-2260

66 Bar-Even, A.*, et al.* (2010) Design and analysis of synthetic carbon fixation pathways. *Proc. Natl. Acad. Sci. USA* 107, 8889-8894

67 Whitney, S.M.*, et al.* (2011) Advancing our understanding and capacity to engineer nature's CO_2-sequestering enzyme, Rubisco. *Plant Physiol.* 155, 27-35

68 Yohn, C.*, et al.* (2011) Sapphire Energy, Inc. Stress-induced lipid trigger. Application no.: 11740278.4. Published source: European Patent Register.

69 Trentacoste, E.M.*, et al.* (2013) Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth. *Proc. Natl. Acad. Sci. USA* 110, 19748-19753

70 Dahl, R.H.*, et al.* (2013) Engineering dynamic pathway regulation using stress-response promoters. *Nat. Biotechnol.* 31, 1039-1046

71 Kilian, O.*, et al.* (2011) High-efficiency homologous recombination in the oil-producing alga *Nannochloropsis* sp. *Proc. Natl. Acad. Sci. USA* 108, 21265-21269

72 Purton, S.*, et al.* (2013) Genetic engineering of algal chloroplasts: progress and prospects. *Russ. J. Plant Physiol.* 60, 491-499

73 Ohresser, M.*, et al.* (1997) Expression of the arylsulphatase reporter gene under the control of the *nit1* promoter in *Chlamydomonas reinhardtii*. *Curr. Genet.* 31, 264-271

74 Quinn, J.M.*, et al.* (2003) Copper response element and Crr1-dependent Ni^{2+}-responsive promoter for induced, reversible gene expression in *Chlamydomonas reinhardtii*. *Eukaryot. Cell* 2, 995-1002

75 Helliwell, K.E.*, et al.* (2014) Unraveling vitamin B_{12}-responsive gene regulation in algae. *Plant Physiol.* 165, 388-397

76 Ramundo, S.*, et al.* (2013) Repression of essential chloroplast genes reveals new signaling pathways and regulatory feedback loops in *Chlamydomonas*. *Plant Cell* 25, 167-186

77 Fang, W.*, et al.* (2012) Transcriptome-wide changes in *Chlamydomonas reinhardtii* gene expression regulated by carbon dioxide and the CO_2-concentrating mechanism regulator CIA5/CCM1. *Plant Cell* 24, 1876-1893

78 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, e43349

79 Day, A. and Goldschmidt-Clermont, M. (2011) The chloroplast transformation toolbox: selectable markers and marker removal. *Plant Biotechnol. J.* 9, 540-553

Highlights

- Microalgae are potential hosts for industrial biosynthesis of valuable compounds.
- Genome sequences of many microalgae are available, but annotation lags behind.
- We propose an integrative approach to improve algal proteins annotation.
- Systems biology modeling of microalgae is crucial to facilitate algal engineering.

Figure 1

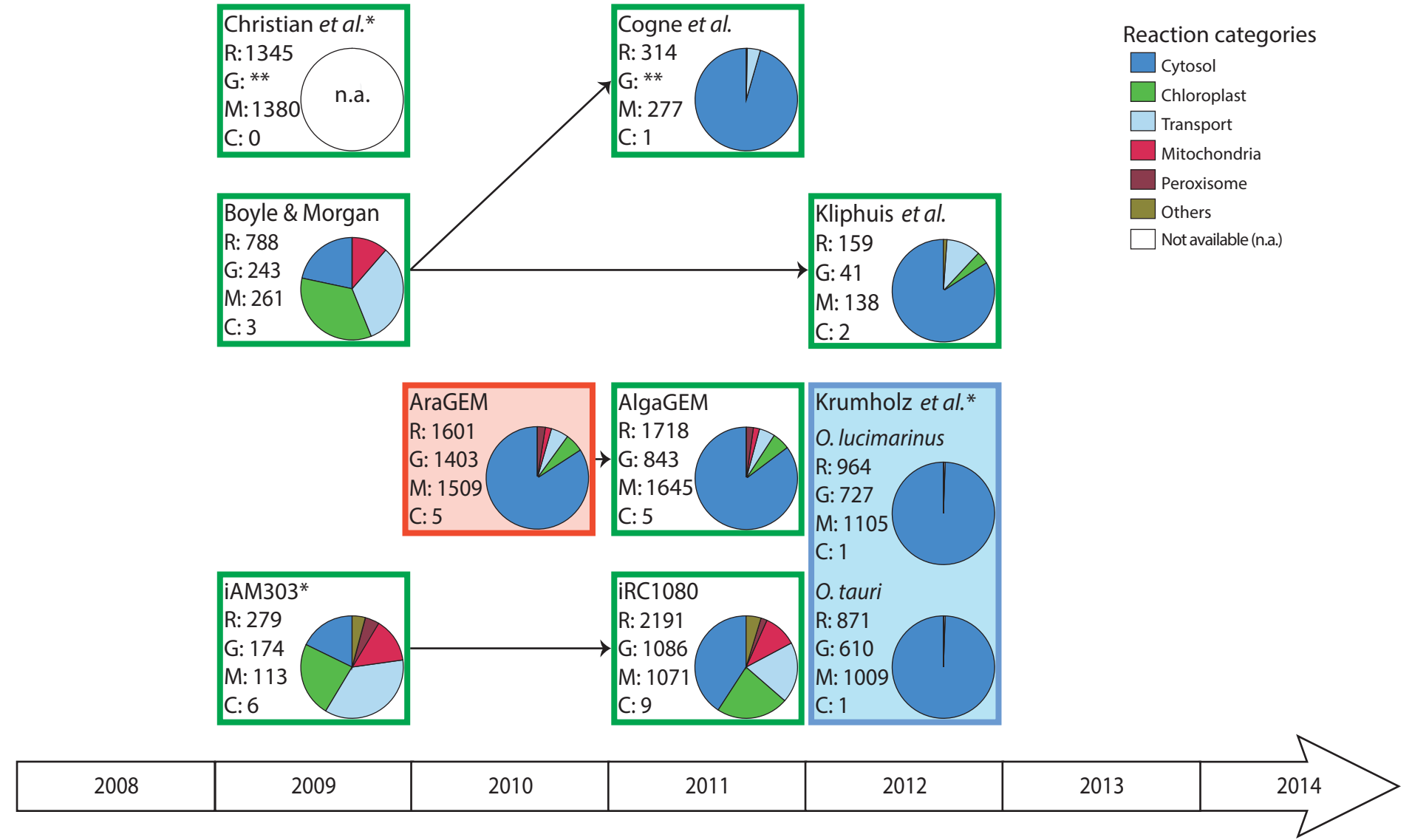
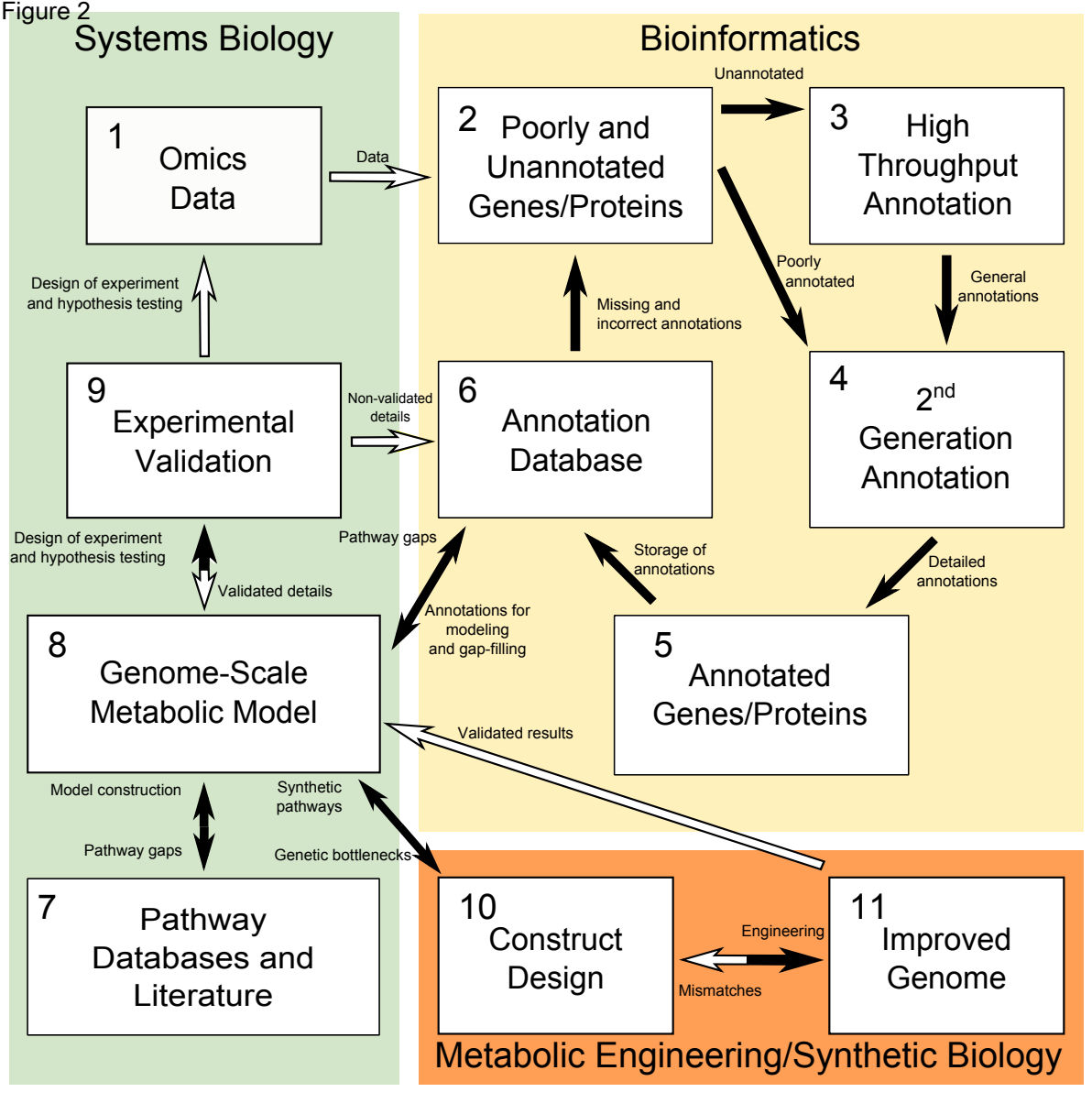

Christian et al.*
R: 1345
G: **
M: 1380
C: 0
n.a.
Cogne et al.
R: 314
G: **
M: 277
C: 1
Boyle & Morgan
R: 788
G: 243
M: 261
C: 3
Kliphuis et al.
R: 159
G: 41
M: 138
C: 2
AraGEM
R: 1601
G: 1403
M: 1509
C: 5
AlgaGEM
R: 1718
G: 843
M: 1645
C: 5
Krumholz et al.*
O. lucimarinus
R: 964
G: 727
M: 1105
C: 1
O. tauri
R: 871
G: 610
M: 1009
C: 1
iAM303*
R: 279
G: 174
M: 113
C: 6
iRC1080
R: 2191
G: 1086
M: 1071
C: 9
Reaction categories
Cytosol
Chloroplast
Transport
Mitochondria
Peroxisome
Others
Not available (n.a.)
2008
2009
2010
2011
2012
2013
2014

Figure 2


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 2nd 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
Pathway gaps
Synthetic pathways
Genetic bottlenecks
7 Pathway Databases and Literature
10 Construct Design
Engineering
Mismatches
11 Improved Genome
Metabolic Engineering/Synthetic Biology

Original Figure File_Fig.1
Click here to download Original Figure File: Figure_1_Adobe Illustrator.ai

Original Figure File
Click here to download Original Figure File: Figure_2_Inkscape.svg