Review

CellPress

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

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 biosynthesis 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

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

0167-7799/
 http://dx.doi.org/10.1016/j.tibtech.2014.10.003

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 for success Suitable for high-throughput analysis	[16]
HMMER	Moderate	Fast	Online/offline	Family-wise alignment method Suitable for high-throughput analysis	[17]
InterProScan	Moderate	Slow	Online/offline	Family-wise alignment method Uses pre-computed protein domains	[72]
FFPred2	High	Slow	Limited online/offline	Algorithms currently trained on non-algal datasets Not suitable for high-throughput analysis	[20,23]
Argot2	High	Moderate	Limited online	Initial selection is dependent on BLAST and HMMER output Additionally predicts compartmentalization User-friendly interface	[15]

[a]For distantly related sequences.

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

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.