Genome-wide metabolic network reconstruction of the picoalga Ostreococcus

Elias W. Krumholz¹, Hong Yang¹, Pamela Weisenhorn¹,², Christopher S. Henry³ and Igor G. L. Libourel¹,*

¹ Department of Plant Biology, 1500 Gortner Avenue, University of Minnesota, Saint Paul, MN 55108, USA
² Department of Ecology, Evolution and Behavior, 1987 Upper Buford Circle, Saint Paul, MN 55108, USA
³ Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA

* To whom correspondence should be addressed. E-mail: libourel@umn.edu

Abstract

The green picoalga Ostreococcus is emerging as a simple plant model organism, and two species, O. lucimarinus and O. tauri, have now been sequenced and annotated manually. To evaluate the completeness of the metabolic annotation of both species, metabolic networks of O. lucimarinus and O. tauri were reconstructed from the KEGG database, thermodynamically constrained, elementally balanced, and functionally evaluated. The draft networks contained extensive gaps and, in the case of O. tauri, no biomass components could be produced due to an incomplete Calvin cycle. To find and remove gaps from the networks, an extensive reference biochemical reaction database was assembled using a stepwise approach that minimized the inclusion of microbial reactions. Gaps were then removed from both Ostreococcus networks using two existing gap-filling methodologies. In the first method, a bottom-up approach, a minimal list of reactions was added to each model to enable the production of all metabolites included in our biomass equation. In the second method, a top-down approach, all reactions in the reference database were added to the target networks and subsequently trimmed away based on the sequence alignment scores of identified orthologues. Because current gap-filling methods do not produce unique solutions, a quality metric that includes a weighting for phylogenetic distance and sequence similarity was developed to distinguish between gap-filling results automatically. The draft O. lucimarinus and O. tauri networks required the addition of 56 and 70 reactions, respectively, in order to produce the same biomass precursor metabolites that were produced by our plant reference database.

Key words: Gap-filling, metabolic network, network reconstruction, Ostreococcus.

Introduction

Due to the large research investments in genome projects and the rapid advancement of sequencing technologies, the number of sequenced genomes is growing exponentially (Koonin et al., 2008; Kyrpides, 2009). These sequences have great potential value, but their use is limited by the amount of time and effort required functionally to annotate a genome. Genes annotated with metabolic reactions are readily interpretable at the biochemical reaction level, but their metabolic function is dependent on which other reactions are present. Flux balance analysis (FBA) (Orth et al., 2010) performs such a functional evaluation and has the capability to evaluate in which metabolic functions a reaction participates. FBA is therefore an excellent technology to evaluate the annotations for metabolic genes (Henry et al., 2011). Before such functional analysis can be performed, all the reactions associated with annotated metabolic genes must first be aggregated into a metabolic network. For prokaryotes, metabolic network reconstruction has become routine and, in many cases, sequence annotation and network reconstruction can be produced in a fully automated fashion (Henry et al., 2010). The quality of such machine annotations is dependent on the ability to take contextual information into consideration during the annotation process. For instance, prokaryote annotation algorithms take
the location of a gene relative to other functionally related genes into account. Eukaryotic genomes have much greater complexity, and the location of genes in eukaryotic genomes is much less informative. Consequently, annotation methods developed for prokaryotes have struggled when applied to plant genomes, requiring that these genomes still be annotated by expert teams (Wang et al., 2011). A metabolic network by itself can potentially provide a wealth of contextual information that is also applicable to eukaryotic systems. Metabolism can be viewed as multifaceted, highly interdependent machinery, containing functionality that is easily computer interpretable. Missing or superfluous reactions in the metabolic network can be readily identified and addressed by modifying the network in such a way that the functional metabolic unit is restored.

Here FBA has been applied on metabolic networks to evaluate the completeness of metabolic annotations for two *Ostreococcus* species. The prevalent marine microalga *Ostreococcus* (Zhu et al., 2005) is an ideal model organism in plant biology due to its simplicity and its phylogenetic position as an early-diverging green plant lineage (Derelle et al., 2006; Palenik et al., 2007). *O. tauri* is the smallest known existing eukaryote (<1 μm), it can be kept in culture and can be genetically transformed (Corellou et al., 2009). *Ostreococcus* is haploid (Corellou et al., 2009; Grimsley et al., 2010) and has a single copy mitochondrion and chloroplast. *Ostreococcus* has been discovered relatively recently (Courties et al., 1994; Chretiennot-Dinet et al., 1995), but its importance is broadly recognized which has resulted in over 150 scientific publications, of which 70 were published in just the last two years. The significance of *Ostreococcus* is further exemplified by the complete genome sequencing of three species, and the resequencing of 15 more. Two of these genome sequences, *O. tauri* and *O. lucimarinus*, have been manually annotated (Derelle et al., 2006; Palenik et al., 2007), setting the stage for metabolic network analysis.

Besides the quality of an organism’s annotation, the ability to reconstruct its metabolic network algorithmically depends on a well-curated biochemical reaction database with gene-to-reaction associations. Gene-to-reaction mappings are organized in an orthology database that associates sequences of individual species with a biochemical reaction and allows for the identification of probable homology between organisms. FBA requires that the reactions included within the metabolic network be balanced at the element level. If a reaction is not elementally balanced, FBA will produce biologically meaningless solutions. For instance, a network that contains an oxygen unbalanced reaction might apply that reaction as part of a cycle consuming all oxygen produced by photosynthesis. One of the best known large ontology databases associated with biochemical reactions is produced by KEGG (Kanehisa et al., 2006). This work makes use of the balanced subset of the KEGG Orthology (KO) database to associate the gene annotations with biochemical reactions.

Flux balance analysis of the reconstructed draft networks reveals network functionality for some pathways, but more importantly a lack of functionality for others. Non-functional pathways can be gap-filled by adding reactions to the network until the demanded network functionality is achieved (Orth and Palsson, 2010). Gap-filling requires a large reference database of reactions that may be used to fill the network gaps. For this purpose, the complete set of balanced reactions in the KEGG KO database was used. The KO database spans all kingdoms of organisms and many of the reactions exist in microbial organisms only, making them unsuitable for the gap-filling of plant networks. To address this potential issue, a layered gap-filling approach was introduced, where the *Ostreococcus* networks were almost exclusively filled with reactions known to exist in the set of plants annotated in the KEGG database. This database of plant reactions has been called the metaplant. The meta-plant database was curated using nested layers of the KEGG database in an attempt to retain the functionality of the complete KO database. Hence, the KO database biomass capability represents the maximum feasible biomass any model based on gene annotations from KEGG can achieve. Using a gap-filling algorithm, this functionality can be added to smaller databases through the addition of a minimal set of reactions from the KO database. The set of all eukaryotic and cyanobacteria annotations was gap-filled using the KO database to produce a reduced database, which was subsequently used to gap-fill the metaplant.

The model systems *Ostreococcus*, *Arabidopsis*, and *Chlamydomonas* represent three clades that provide the full scope of green plant-specific genes: ‘the green cut’ (Merchant et al., 2007). Curated genome-wide metabolic networks for *Arabidopsis* (Poolman et al., 2009; Dal’Molin et al., 2010) and *Chlamydomonas* (May et al., 2008, 2009; Christian et al., 2009) already exist, and this work presents and compares the metabolic reconstructions of two *Ostreococcus* species.

Materials and methods

Functional gene annotations were collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology (KO) database on 28 April 2011. This database contains mappings between the KEGG KO identifiers, organism-specific genes, predicted enzyme functionalities (EC numbers), and KEGG reactions. In addition, the KEGG reaction, compound, and enzyme databases were downloaded on the same date in flat file format from the KEGG FTP website. The databases were loaded into Matlab® (The MathWorks, Natick, MA) structures and organized according to the flat file field names.

To generate an SBML model from a KEGG genome, metabolic genes must be linked to metabolic reactions, but KEGG does not provide such a mapping. Instead, genes annotated with a functional role are assigned a KEGG orthology identifier (KO number). Most KO entries point directly to a set of reactions, all of which were included. If this was not the case, a KO entry often pointed to an Enzyme Classification (EC) identifier, in which case all reactions associated with the enzyme activity were added to the KO structure. The complete database mapping structure is shown in Supplementary Fig. S1 at JXB online.

The database structure was then reorganized to be rooted at the reaction level. Unique reaction identifiers were annotated with (potentially multiple) KO identifiers, EC numbers, and genes associated with these KO identifiers and EC numbers. After removal of unbalanced or incomplete reactions, SBML models (level 2, version 4) were generated using the System Biology
Network reconstruction of Ostreococcus

toolbox (Keating et al., 2006). The SBML reaction field was populated with the organism-specific subset of the reaction database. Compounds were pulled from this reaction set and added to the SBML species field. Multi-organism models were generated by creating a union of organism-specific reaction databases. Each model was supplemented with a list of spontaneous reactions (see Supplementary Table S1 at JXB online). SBML models were subsequently converted to COBRA compliant format to access COBRA toolbox functionality. COBRA toolbox v2.0 (Schellenberger et al., 2011) was downloaded from the openCOBRA project at sourceforge.net. (http://opencobra.sourceforge.net/). The generated models were un compartmentalized.

Gene comparisons

Genomes of organisms were downloaded in FASTA format from the KEGG database on 28 April 2011. A best gene match between genes in O. tauri and O. lucimarinus and genes in the union of KEGG plant genomes (meta-plant) was found using the Smith-Waterman algorithm (Smith et al., 1981) performed on a TimeLogix® DeCypher® (Active Motif Inc., Carlsbad, CA) gene comparison server. The union of plant genes for each reaction present in the KEGG database was used as a query sequence and the gene models of O. tauri and O. lucimarinus were used as the databases to search against. In this way, a mapping between each metabolic gene in the meta-plant genome and the best matching gene from both O. tauri and O. lucimarinus was created. Once the meta-plant model was complete, each reaction present in the meta-plant model was associated with the best scoring gene comparison. This method allowed every reaction in a large database to be annotated with a specific gene from an organism of interest and a corresponding gene from a plant database regardless of gaps in previous annotations or poor sequence similarity between available genes annotated with a particular reaction.

Elemental balance of reactions

The KEGG orthology database (28 April 2011) contained mappings to 4523 reactions. The elementary mass balance of each reaction was tested using a custom-developed Matlab routine. To prevent stoichiometric matrix errors, reactions that contained the same metabolite as substrate and product were removed (see Supplementary Table S2 at JXB online). The results were verified with the elementary balancing functionality of the Cobra toolbox and no discrepancies were found. Generic compound equations containing (n) or R-groups were substituted with a large prime number for n or R-groups were substituted. These reactions were removed, with the exception of a small set of reactions that were manually balanced to retain the ability to reach five biomass precursor metabolites (see Supplementary Table S3 at JXB online).

Reversibility index for reactions

The reversibility of reactions was determined using the free energy calculations for reactions based on a group theory approach (Mavrovouniotis, 1991; Forsythe et al., 1997) and Jankowski et al. (2008), which was further refined by the Milo laboratory (Flamholz et al., 2011). Elad Noor (Milo laboratory) kindly provided a custom reaction list adjusted to pH 7.5 and an ionic strength of 0.3 upon request. The reversibility index was generated according to the metric developed by Noor et al. (E Noor et al., unpublished data). Default metabolite concentrations were assumed to be 100 μM and allowed to vary between 3 μM and 3 mM which corresponds to an index cut-off value of 1000. Using these constraints, approximately half of the reactions in the KO database were considered irreversible. Reversibility information was included in the first two kinetic parameters of each reaction following the COBRA format. A reversible reaction was added as a chemical description of photosynthesis to allow the model to intake energy. This reaction is listed as R99999 (equation: 2H₂O+4 oxidized ferredoxin→4H⁺+O₂+4 reduced ferredoxin).

Biomass production evaluation

Network functionality was investigated using the flux balance analysis (FBA) capability of COBRA. For this purpose, a target biomass component list was compiled from a literature survey of published plant genomes (see Supplementary Tables S4 and S5 at JXB online). Of this list, only the compounds that could be mapped to KEGG identifiers were included in our biomass definition (see Supplementary Table S4 at JXB online). Missing functionality in target (reference) networks was addressed by detecting and gap-filling missing reactions with the SMILEY algorithm (Reed et al., 2006; Orth et al., 2010). The MILP objective was set to optimize for a maximum number of biomass components by adding reactions or allowing metabolites to escape the network through transporters. The meta-plant databases was generated stepwise by first producing a complete KO reaction model that was gap-filled with hydrogen-unbalanced reactions to increase the number of reachable biomass precursor metabolites by five. These unbalanced reactions were manually corrected and added to the network (see Supplementary Table S3 at JXB online). Reference models of decreasing taxonomic diversity were gap-filled stepwise with the networks of one level higher divergence until the generic plant model (meta-plant) was produced. Each gap-filling step ran for 20 iterations to find different reaction combinations to enable the production of all biomass metabolites. The reaction list of a solution requiring the fewest reactions was added to gap-fill the reference models.

Phylogenetic reconstruction

The phylogenetic distance among all KEGG plant taxa was inferred from publicly available, fully sequenced and annotated genomes. Phylogenetic distance between species was estimated from six nuclear protein-coding genes: isoleucyl-tRNA synthetase, arginyl-tRNA synthetase, ribosomal protein L14, ribosomal protein S7, DNA-directed RNA polymerase alpha subunit, and DNA-directed RNA polymerase beta subunit (see Supplementary Table S6 at JXB online). These genes were previously identified by Ciccarelli et al. (2006) as useful for reconstructing phylogenies among widely divergent taxa.

Gene sequences were downloaded from the KEGG Genome database using the KEGG ID as a search string. Many genes have variable copy number within and among taxa; therefore, single consensus sequences were generated for genes with multiple copies using Clustal X (Larkin et al., 2007) by aligning the copies and generating a single, consensus sequence (see Supplementary Table S6 at JXB online). Gene sequences were then aligned across taxa using Clustal X with default parameters. The software package jModelTest 0.1 (Posada, 2008) was used to select the best fitting-model of nucleotide evolution for each gene individually using the Akaike information criterion (AIC). The generalized time reversible (GTR) model with branch-specific evolutionary rates following a gamma distribution (GTR+G) and independent frequencies for each nucleotide (GTR+I+G) was chosen for isoleucyl-tRNA synthetase (K01870), while the GTR+G model with equal nucleotide frequencies was chosen for all other genes. Genes were concatenated by hand. A maximum-likelihood (ML) tree was then inferred from the concatenated and partitioned genetic data set using Garli 2.0 (Zwickl, 2006). Models parameters estimated using jModelTest were used for each gene, and a cladogram based on current systematic knowledge (G Weiblen, University of Minnesota, personal communication) was enforced as a constraint to ensure accurate topology. All other parameters were left at the default values. The default termination criteria were used to determine when the run was complete. A Newick string with distances was converted to a distance matrix using the ape package for R 2.10.1 (R Core Development Team, 2009) (see Supplementary Table S7 at JXB online).
Table 1. Scale and functionality of networks
The eukaryote–cyanobacteria model was refined using bottom-up gap-filling from the KEGG Orthology (KO) database. The meta-plant model was subsequently gap-filled with the gap-filled eukaryote–cyanobacteria model using the bottom-up method. Finally, the Ostreococcus networks were gap-filled with the gap-filled meta-plant model using the top-down approach. In all gap-filling instances, more reactions were added to the model than metabolites indicating an increase in overall network connectivity. All producible biomass functionality was transferred from the KO model to the eukaryote–cyanobacteria model. However, the meta-plant model was only able to produce 48 biomass components, one less than the KO and eukaryote–cyanobacteria models. Both Ostreococcus models were able to produce all 48 biomass components from the meta-plant model after gap-filling even though no biomass components were producible in O. tauri prior to gap-filling. Gap-filling increased the percentage of reactions with feasible fluxes in both Ostreococcus species, suggesting a substantial improvement in network connectivity as a result of gap-filling. Feasible (%)=100×number of feasible fluxes/(number of feasible fluxes+number of non-feasible fluxes).

<table>
<thead>
<tr>
<th>Gap-filling</th>
<th>No of reactions</th>
<th>No. of metabolites</th>
<th>Producible biomass</th>
<th>Feasible (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG orthology</td>
<td>None</td>
<td>3937</td>
<td>3582</td>
<td>49</td>
</tr>
<tr>
<td>Eukaryote–cyanobacteria</td>
<td>None</td>
<td>2970</td>
<td>2826</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Bottom-up</td>
<td>2974</td>
<td>2827</td>
<td>49</td>
</tr>
<tr>
<td>Meta-plant</td>
<td>None</td>
<td>2060</td>
<td>2153</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Bottom-up</td>
<td>2068</td>
<td>2154</td>
<td>48</td>
</tr>
<tr>
<td>O. lucimarinus</td>
<td>None</td>
<td>908</td>
<td>1076</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Top-down</td>
<td>964</td>
<td>1100</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>36.47</td>
</tr>
<tr>
<td>O. tauri</td>
<td>None</td>
<td>801</td>
<td>971</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Top-down</td>
<td>871</td>
<td>1014</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>48.92</td>
</tr>
</tbody>
</table>

Network visualization
Cytoscape v2.8.1 (Smoot et al., 2011) was used to generate metabolic network visualizations from SBML level 2 version 1 files. The advanced network merge plugin was used to create a difference network for O. lucimarinus against O. tauri and O. tauri against O. lucimarinus. The difference networks were combined with the union function to generate a complete difference network. The network was rendered with the VizMapper function using the yFiles ‘organic’ layout algorithm. Organism-specific reactions were identified in VizMapper by storing identifier strings in the sbml ‘name’ field of reactions (Smoot et al., 2011).

Results and discussion
Functional network analysis
FBA was used to investigate the ability of reconstructed networks to produce biomass components. The unedited O. lucimarinus and O. tauri networks were able to produce 18 and 0 biomass components, respectively (Table 1). Limited network functionality of draft networks reconstructed from genomic databases is not uncommon, and is predominantly caused by the presence of gaps in the network. Network gaps arise from missed annotations, and in the case of KEGG, disconnects between generic and specific definitions of the same metabolites. Other gaps arise from the removal of unbalanced equations during the network reconstruction process. The difference in network annotation between O. tauri and O. lucimarinus also suggest that O. tauri was annotated more conservatively. An example of the conservative annotation of O. tauri was the lack of Calvin cycle capability [ribulose-5-phosphate-3-epimerase (EC: 5.1.3.1)] was not included in the draft network].

Gap-filling
Two complementary approaches to gap-filling of metabolic networks exist in the literature. Bottom-up gap-filling is based on a mixed integer optimization routine usually aiming to add a minimal number of reactions to a network (Reed et al., 2006). This method can distinguish between different classes of reactions either by adding reactions in a preferred order, or by associating different weights with the different reaction classes (Henry et al., 2010). The bottom-up gap-filling method is the most commonly applied approach, and it was used for gap-filling all the reference databases described below.

A top-down approach to gap-filling, pioneered by (Christian et al., 2009), adds the complete gap-filling reaction database to a draft network followed by the iterative removal of the added reactions. This removal process is continued until no more added reactions can be removed without losing biomass production capability. Both the top-down and bottom-up methods are iterative approaches that do not have unique solutions. Candidate gap-filling reactions from species closely related to the target species are more likely to feature in the target species’ network. By gap-filling reference networks of decreasing taxonomic diversity, a layered approach to gap-filling was used that takes advantage of this quality. (Fig. 1) With this in mind, gap-filling of the Ostreococcus networks was performed with a list of just the reactions that KEGG has associated with the 17 plant genomes in their database. This meta-plant network was not free of gaps itself, and has been filled with the combined reactions of eukaryotes and cyanobacteria. Similarly, the cyano–eukaryote model was gap-filled using the complete KEGG ontology reaction list. The complete network database was able to produce 44 out of the 94 defined biomass precursor metabolites, which increased to 49 after adding 12 previously H-unbalanced reactions (see Supplementary Table S3 at JXB online). After the consecutive gap-filling of increasingly small networks, the smaller networks approximated the biomass production capability of the ontology dataset (Table 1). This biomass component number fell well short of the target list of over 90
network was associated with all sequences that were anno-
ticated. Sequences are compared.

To calculate this score, each reaction within the meta-plant network was associated with all sequences that were annotated with that reaction. Every gene associated with any reaction was then compared with all genes in the O. tauri and O. lucimarinus genomes using the Smith–Waterman algorithm. The Smith–Waterman e-score is a likelihood score applicable to a set of translated amino acids, which was weighted with the phylogenetic distance between respective species: \( \omega = \sum -p_i \log(e_i) \), where \( e \) is the Smith–Waterman e-score and \( p \) is the normalized phylogenetic distance score. Each sequence comparison thus yields a quality-score, and the best quality score, \( \omega_i \), for each reaction is assigned to that reaction in the meta-plant network: \( \omega = \min(\omega_j) \), where \( j \) is the number of reference sequences associated with the reference reaction. The \( \epsilon \)-value was capped at a value of one to prevent sign inversion. The domain of \( \omega \) is from 0 (perfect match) to \( \infty \) (very poor match), i.e. if a reaction had been annotated for the target organism, the value for that reaction would be zero (zero over minus infinity). A lack of a close sequence comparison would result in a value proportional to the phylogenetic distance and inversely proportional to the \( \epsilon \)-value. This quality metric allows for the rapid discrimination between gap-filling solutions (Fig. 2). Phylogenetic relationships among plant species included in the KEGG database was inferred using a maximum likelihood approach, see the Materials and methods for a detailed explanation.

**Gap-filling of the Ostreococcus networks**

The Ostreococcus networks were gap-filled using the meta-plant network as the reference reaction database. To enable the production of all 48 biomass components produced by the reference database, the O. tauri and O. lucimarinus networks gained 70 and 56 reactions respectively (Table 1). The iterative nature of the employed gap-filling methods resulted in multiple solutions. The results presented in Table 1, were selected based on their overall quality scores. For the O. lucimarinus network, alternative solutions did exist that included one less reaction, but these solutions had a lower quality score. Note that the number of reactions capable of carrying flux dramatically increased upon gap-filling, demonstrating the significantly improved connectivity of the networks. The included reactions with the worst sequence comparison (O. lucimarinus: \( \epsilon = 0.63 \), O. tauri: \( \epsilon = 0.21 \)) indicates that both networks were filled with at least one highly unlikely reaction. This is an unfortunate reflection of the lack of completeness of the gap-filling reaction database.

The bottom-up algorithm enabled the production of only a subset of the 48 biomass metabolites. A comparison of the two methods for the O. tauri network is shown in Fig. 3. For a valid comparison, the top-down method was made to fill the O. tauri network for the 36 biomass metabolites that the bottom-up method found. In this direct comparison between the best gap-filling solutions for the two methods, the bottom-up method used seven fewer reactions than the top-down approach. However, the top-down approach had a better quality score, indicating a more realistic gap-filling solution. Finally, poor estimates of the imposed thermodynamic constraints could have led to incorrect reversibility constraints, causing unrealistic pathway shunts to accomplish the required biomass capability. However, if thermodynamic constraints had not been imposed, flux balance analysis could have found solutions that make use of thermodynamically infeasible pathways. That the thermodynamic constraints were active is readily demonstrated: FBA of the unconstrained O. lucimarinus network yielded a biomass flux of 2.079, compared to 0.275 for the constrained version, underlining the importance of accurate thermodynamic constraints. Other previously published studies also emphasize the importance of thermodynamic constraints to the accuracy of FBA models (Faria et al., 2010).
Fig. 2. Network quality metric. Draft networks can be gap-filled with the fewest number of added reactions (1A). Alternatively, reactions can be weighted for their likelihood to exist in a target organism (1B) by considering the phylogenetic proximity (2A) and sequence orthology (2B) of best matching sequences. Both factors were included in a quality metric (ω) and associated with each reaction (3). This allowed the top-down gap-filling algorithm to preferentially include reactions with low (good) quality scores and was used to compare gap-filling results. All genes annotated with a particular reaction in the reaction database have a best match with a gene in both O. tauri and O. lucimarinus. A gap-filling solution with the lowest sum of quality metrics for all reactions in a network was considered the best solution.

Fig. 3. Comparison of top-down and bottom-up gap-filling. Fair comparison between top-down and bottom-up gap-filling requires the biomass targets to be identical. Because the bottom-up algorithm was only able to produce 36 biomass components, these same 36 components were used as a target for the top-down method. Top-down gap-filling added a total of 57 reactions to the draft O. tauri metabolic network of 801 reactions. The bottom-up algorithm added 50 reactions, 44 of which were also present in the top-down solution. Although the bottom-up algorithm included seven fewer reactions than the top-down algorithm, the combined quality score for the 50 added reactions was 6.204 whereas the top-down method scored 6.044 for 57 reactions. Comparison of gap-filling results Before gap-filling 25.8% of the reactions in the O. tauri network were capable of carrying flux (Feasible) using the available uptake metabolites while allowing all other metabolites to export. After bottom-up gap-filling 48% of the reactions were capable of carrying flux. After top-down gap-filling 48.9% of reactions could carry flux. When the top-down approach was made to produce the entire meta-plant biomass target of 48 reactions, 70 reactions were added with a combined quality score of 7.252 and 49.8% of the reactions in the network could carry flux. The network quality score is the sum of all quality scores of the reactions included in the network, and the maximum ω-value indicates the worst reaction quality score of the included reactions.
Network reconstruction of Ostreococcus

Fig. 4. Reactions added by gap-filling to O. tauri and O. lucimarinus. The draft networks of O. tauri and O. lucimarinus consisted of 801 and 908 reactions respectively, and contained an overlap of 765 reactions. The gap-filling process added a total of 70 reactions to O. tauri and 56 reactions to O. lucimarinus. O. tauri and O. lucimarinus shared 53 gap-filled reactions, which were present in neither draft network. O. lucimarinus donated 13 reactions to O. tauri during the gap-filling process, but O. tauri only donated one reaction to O. lucimarinus. Only six gap-filled reactions were unique to a single network, four were added to O. tauri and two were added to O. lucimarinus. Despite the large amount of shared reactions added during the gap-filling process, the networks retained many of their unique reactions (reactions present in only O. tauri or O. lucimarinus): the gap-filled O. tauri contained 39 unique reactions, and the gap-filled O. lucimarinus contained 132 unique reactions.

Network comparison

A large share of the reactions added during the gap-filling process were added to both networks (Fig. 4). Only a single reaction originally annotated for O. tauri was added to the O. lucimarinus network. Conversely, 13 reactions originally annotated for O. lucimarinus were added to the O. tauri network. Two and four reactions were added to the networks of O. lucimarinus and O. tauri, respectively, that were unique to the networks. Network changes resulting from gap-filling are shown in Fig. 5.

The differences between the draft Ostreococcus networks were visualized (Fig. 6) by calculating a difference network that only shows reactions exclusively present in only one network (logical XOR). The same difference network was generated after gap-filling, and the changes between the difference networks, which resulted from gap-filling, are shown in a third panel. Because only the connected differences between the networks are shown, the connectivity of the difference network shows the alternative routes in central metabolism utilized by the Ostreococcus species connecting ribose and glyoxylate metabolism. The differentially added reactions show increased divergence between the two Ostreococcus networks, but a much larger number of reactions disappeared after gap-filling, illustrating the converging effect on the networks resulting from gap-filling.

Comparison to existing reconstructions

Due to the relatively recent discovery of Ostreococcus, little biochemical data are readily available. Consequently, the presented genome-scale reconstructions were exclusively based on genomic orthology. In comparison, for the established model green alga Chlamydomonas, at least three large-scale metabolic reconstructions exist (Boyle and Morgan, 2009; Chang et al., 2011). One of these models (Boyle and Morgan, 2009) is a detailed manual reconstruction that focuses on a comparison of predictions for heterotroph, autotroph, and mixotroph growth conditions. This reconstruction was not genome-wide (458 metabolites and 484 metabolic reactions), but it was compartmentalized and contained an extensive description of photosynthesis to investigate linear and cyclic electron transport. The first genome-scale model of Chlamydomonas was produced to introduce the bottom-up gap-filling algorithm (Christian et al., 2009). This model was uncompartmentalized, and compares most closely in scope and approach to the Ostreococcus models reconstructed in this paper. Recently, a second genome-wide Chlamydomonas reconstruction appeared (Chang et al., 2011) that includes cellular compartmentation. This work also addresses the role of light in algal metabolism and is the most sophisticated algal model to date. The network has roughly double the number of reactions in the Ostreococcus models, with the Chlamydomonas model having 2019 reactions and 1069 metabolites.
Conclusions

The layered construction of the meta-plant reference database prevented incorporation of microbial reactions where possible. However, the bottom-up gap filling algorithm was unable to maintain the full biomass capability whilst gap-filling networks of reduced taxonomy (see Supplementary Tables S8 and S9 at JXB online). This limitation was also encountered during gap-filling of the Ostreococcus networks using the meta-plant network. By contrast, the trimming algorithm was able to retain all biomass functionality albeit at the cost of requiring more reactions.

The complete reference database was able to produce just over half of the target biomass metabolites. This biomass list is more extensive and varied than most models in the literature, but some common biomass components could not be produced (see Supplementary Table S9 at JXB online). This may reflect the limited list of balanced reactions contained within the KEGG database. Due to the limited size of this reference database, the Ostreococcus networks presented should not be regarded as definitive. The development of an exhaustive, open source and curated biochemical reaction list specific to plants should therefore be a priority in the development of plant-model reconstruction technology.

Although the quality of the reconstructed Ostreococcus networks is not on a par with carefully manually curated networks, the reconstruction process highlighted the ability to evaluate the completeness of the genome annotation for the Ostreococcus species. The large number of reactions that needed to be added for the evaluated biomass components suggests that the O. tauri genome annotation in particular is lacking a substantial number of enzyme annotations. Comparison of the two metabolic network reconstructions suggested that O. tauri had been annotated more conservatively than O. lucimarinus. The difference between the two

Fig. 6. O. tauri compared with O. lucimarinus before and after gap-filling. (A, B) Reactions present only in O. tauri or O. lucimarinus before and after gap-filling following binary XOR logic. Reactions present only in O. tauri are shown as light-grey nodes and reactions present only in O. lucimarinus are shown as dark-grey nodes, metabolites are represented by small black nodes. (C) The third network shows the differences between the O. tauri XOR O. lucimarinus before and after gap-filling, including the EC numbers of the selected reactions. The networks show the largest connected component of the XOR graphs in the union of the before and after conditions, and are thus a subsets of the total XOR networks between O. tauri and O. lucimarinus. Reactions represented with squares were removed from the O. tauri XOR O. lucimarinus network during the gap-filling process by adding the corresponding reaction to the other species’ metabolic network. These reactions represent functionality that converged as a result of gap-filling. Stars indicate new additions to the O. tauri XOR O. lucimarinus network as a result of gap-filling. Star reactions were required by only one organism during the gap-filling process and represent diverged functionality. The only star reaction for O. lucimarinus shown here is EC 2.6.1.44 alanine–glyoxylate transaminase. This reaction converts glyoxylate (3) and L-alanine (2) into pyruvate (1) and glycine (4) and is not present in O. tauri. O. tauri also diverged with four unique reactions, one of which (1.4.3.3, D-amino-acid oxidase) involved the interconversion of glyoxylate (3), hydrogen peroxide (5), and ammonia (6) to glycine (4), O2 (not shown), and H2O (not shown). The converged reactions 2.4.2.10 and 4.1.1.23 demonstrate the gap-filling of a missing reaction in O. tauri by incorporating reactions from O. lucimarinus. Reaction 2.4.2.10 converts oridine 5’-phosphate (8) and diphosphate (not shown) into orotate (10) and 5-phospho-alpha-D-ribose 1-diphosphate (11). Reaction 4.1.1.23 converts orotide 5’-phosphate (8), into uridine monophosphate (UMP) (7) and CO2 (9). EC number key: 1.5.1.5 methylenetetrahydrofolate dehydrogenase; 3.5.4.9 methenyltetrahydrofolate cyclohydrolase; 6.3.4.3 formate-tetrahydrofolate ligase; 2.6.1.44 alanine-glyoxylate transaminase; 1.7.7.1 ferredoxin-nitrite reductase; 1.4.3.3 D-amino-acid oxidase; 5.4.99.5 chorismate mutase; 2.4.2.10 orotate phosphoribosyltransferase; 4.1.1.23 orotidine-5’-phosphate decarboxylase; 3.2.2.1 purine nucleosidase; 2.7.1.15 ribokinase.
networks decreased somewhat after gap-filling, suggesting that the difference network of the two species was partly the consequence of the under annotation of \textit{O. tauri}.

Bottom-up and top-down gap-filling approaches are both iterative methods resulting in many solutions. The ability to rapidly evaluate the quality of the gap-filling attempt is essential if many iterations are run or if the network contains many gaps. The \textit{Ostreococcus} networks contained many such gaps, and the introduced measure for network quality provided a valuable tool to discriminate between the many gap-filling solutions automatically. The inclusion of the phylogenetic distance for reactions enriched the networks with reactions of closely related species, and thus the likelihood of these reactions existing within the actual metabolic networks. Recognition of realistic gap-filling solutions and this first network-wide functional comparison between the \textit{Ostreococcus} species will help guide the comprehensive biochemical characterization of \textit{Ostreococcus}.

**Supplementary data**

Supplementary data can be found at \textit{JXB} online.

These data include a complete description of the reference databases, the biomass objective used, and references for each metabolite included in the biomass objective. Metabolic network reconstructions of the \textit{Ostreococcus} species are available in COBRA compatible SBML format.

**Supplementary Fig. S1.** The database mapping structure.

**Supplementary Table S1.** Spontaneous reactions.

**Supplementary Table S2.** Elementally balanced reactions.

**Supplementary Table S3.** Correct H imbalanced reactions list.

**Supplementary Table S4.** Target biomass components in KEGG Orthology database.

**Supplementary Table S5.** Target biomass component not in the balanced KEGG Orthology database.

**Supplementary Table S6.** Nuclear protein: coding genes list.

**Supplementary Table S7.** Phylogenetic distances for \textit{Ostreococcus}.

**Supplementary Table S8.** Reaction list.

**Supplementary Table S9.** Biomass list.

**Acknowledgements**

The authors wish to express their gratitude to George Weiblen (university of Minnesota) for supplying the phylogenetic relationship between plant clades and aiding in the interpretation of the phylogenetic distance estimates. We are grateful for the assistance from Zhengjin Tu for setting up high throughput sequence comparison, and the Minnesota Supercomputing Institute for providing the computational infrastructure to make this work possible. Above all, we value the help we received from Ron Milo, Avi Flamholz, and Elad Noor (Weizmann Institute) in determining the flux directionality based on group contribution calculations. HY and IGLL are supported by NSF MCB award 1042335.

**References**


