Current Biology *16*, 2320–2325, December 5, 2006 ©2006 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2006.09.063

Report

Cyanobacterial Contribution to Algal Nuclear Genomes Is Primarily Limited to Plastid Functions

Adrian Reyes-Prieto,[1] Jeremiah D. Hackett,[1,3] Marcelo B. Soares,[2] Maria F. Bonaldo,[2] and Debashish Bhattacharya[1,*]
[1] Department of Biological Sciences and
Roy J. Carver Center for Comparative Genomics
University of Iowa
446 Biology Building
Iowa City, Iowa 52242
[2] Northwestern University
Children's Memorial Research Center
Chicago, Illinois 60614

Summary

A single cyanobacterial primary endosymbiosis that occurred approximately 1.5 billion years ago [1–3] is believed to have given rise to the plastid in the common ancestor of the Plantae or Archaeplastida—the eukaryotic supergroup comprising red, green (including land plants), and glaucophyte algae [4–8]. Critical to plastid establishment was the transfer of endosymbiont genes to the host nucleus (i.e., endosymbiotic gene transfer [EGT]) [9, 10]. It has been postulated that plastid-derived EGT played a significant role in plant nuclear-genome evolution, with 18% (or 4,500) of all nuclear genes in *Arabidopsis thaliana* having a cyanobacterial origin with about one-half of these recruited for nonplastid functions [11]. Here, we determine whether the level of cyanobacterial gene recruitment proposed for *Arabidopsis* is of the same magnitude in the algal sisters of plants by analyzing expressed-sequence tag (EST) data from the glaucophyte alga *Cyanophora paradoxa*. Bioinformatic analysis of 3,576 *Cyanophora* nuclear genes shows that 10.8% of these with significant database hits are of cyanobacterial origin and one-ninth of these have nonplastid functions. Our data indicate that unlike plants, early-diverging algal groups appear to retain a smaller number of endosymbiont genes in their nucleus, with only a minor proportion of these recruited for nonplastid functions.

Results and Discussion

Several characteristics make *Cyanophora* an ideal Plantae model to study ancient plastid-derived endosymbiotic gene transfer (EGT). This alga retains ancestral cyanobacterial features in its plastid (often referred to as the cyanelle), such as peptidoglycan between the two organelle membranes, concentric (unstacked) thylakoids, and carboxysomes. The host cell is a mesophilic obligate autotroph with a "typical" genome size for free-living protists (circa 140 Mbp [12]), suggesting that *Cyanophora* has not undergone a recent genome reduction because of a parasitic life-style or extremophily. Although glaucophytes are relatively depauperate with respect to taxon richness (only eight genera with 23 species are known [13]) in comparison to their red and green algal (plant) sisters, because of their ancestral plastid features they are often considered to be "living fossils" of the algal world [14].

To assess plastid-derived EGT, we generated 11,176 expressed-sequence tags (ESTs; 3′ single-pass reads) from a normalized *Cyanophora* cDNA library. The exponentially dividing culture was grown under nutrient-replete light conditions. Assembly and clustering of these data resulted in 9,714 high-quality sequences and 3,576 unique genes. Using BLAST, we found significant hits in GenBank (E value cutoff $< 10^{-10}$) to annotated or conserved hypothetical proteins for 1,226 *Cyanophora* genes. To determine whether there was a sampling bias in our approach to EST collection in *Cyanophora* with respect to the distribution of the encoded proteins in different cellular compartments, we placed the 1,226 genes among gene ontology (GO) categories [15]. A total of 446 *Cyanophora* genes could be assigned to the cell, intracellular, and cytoplasm subcategories within the cellular-component category, and these results were compared to equivalent analyses using *Arabidopsis* (2,843/32,719 of total genes assigned [16]), the diatom *Thalassiosira pseudonana* (331/5,002 of total genes assigned [17]), and the green alga *Chlamydomonas reinhardtii* (465/6,627 of total genes assigned [18]). The results of this analysis suggest that our EST collection does not significantly over- or undersample particular cellular-component categories (Figure 1). Therefore, given the number of predicted genes in other free-living, mesophilic algae (e.g., circa 11,242 genes in *Thalassiosira* [19] and circa 15,200 genes in *Chlamydomonas* [18]), the unigene set in *Cyanophora* (3,576 sequences) corresponds, assuming a total gene number between 12,000 and 15,000 in this species, to a representative collection of 24%–30% of the gene repertoire.

To identify genes of cyanobacterial origin in *Cyanophora*, we performed the BLAST search with the 1,226 genes against a local database of all available (17 at the time of writing) complete cyanobacterial genomes (Table S1 in the Supplemental Data available online). This analysis returned 603 (49%) glaucophyte genes with a significant hit to at least one cyanobacterium. This gene set was analyzed in a phylogenomic approach [20] by using a local reference database (see Table S1) of the 17 cyanobacterial, eight diverse eubacterial, two archaeal, five algal/plant (i.e., Plantae), three animal, and three fungal genomes. We examined the 546 resulting neighbor-joining protein-bootstrap trees (Poisson-corrected distance matrices) to identify the topologies that demonstrated weak to moderate (> 60%) or strong bootstrap support for the grouping of *Cyanophora* with one or more cyanobacterial species or included

*Correspondence: debashi-bhattacharya@uiowa.edu
[3] Present address: Woods Hole Oceanographic Institution, Biology Department, MS#32, Woods Hole, Massachusetts 02543.

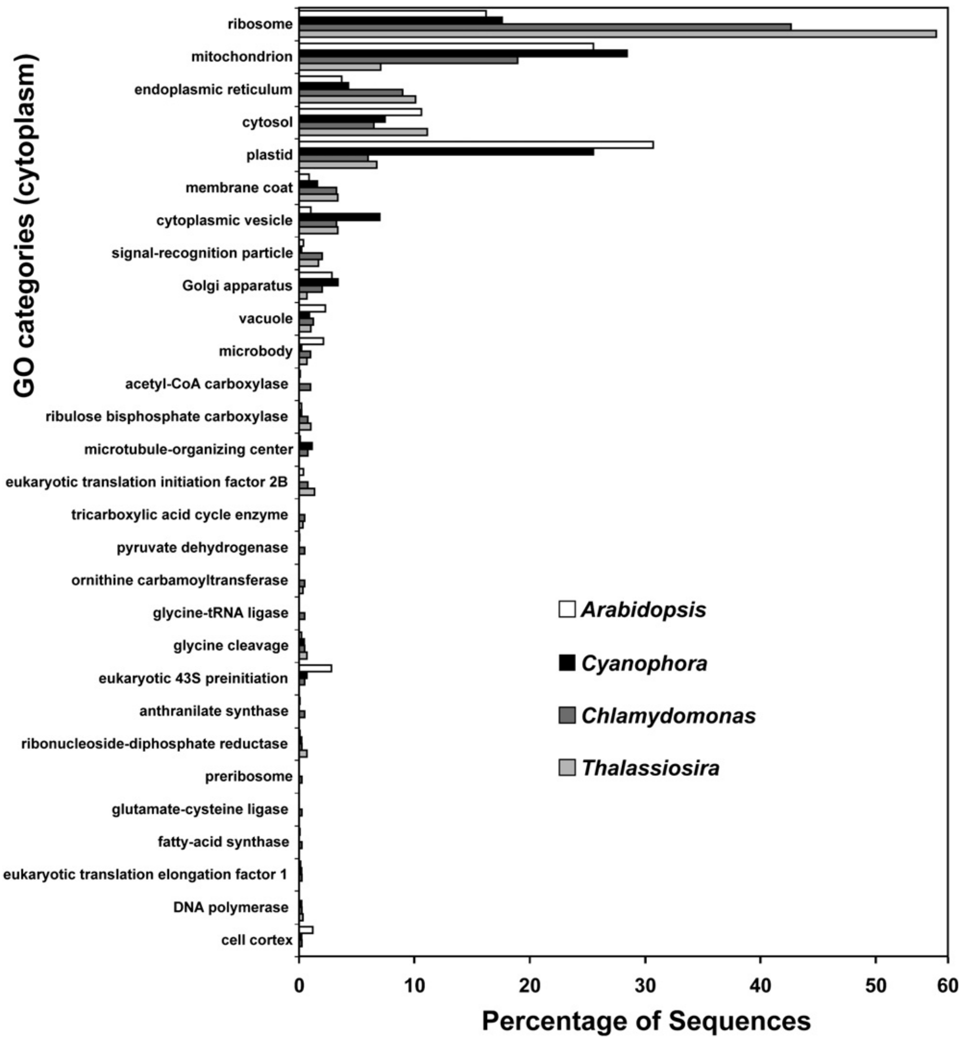

Figure 1. The Distribution of *Cyanophora* Genes in the Cytoplasm Subcategory of the Cellular-Component Ontology Category

The distribution of 1,226 annotated *Cyanophora* genes in GO categories is compared with an equivalent analysis of genome data from *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, and *Thalassiosira pseudonana*. The comparable relative proportions of genes in each category in these photosynthetic eukaryotes suggest the absence of an obvious sampling bias in the *Cyanophora* EST data.

Cyanophora with other members of the Plantae (i.e., *Arabidopsis*, *Oryza*, *Chlamydomonas*, *Cyanidioschyzon*, *Galdieria*). The remaining 57 genes did not return trees when we used the default BLAST and multiple-alignment values that were implemented in the phylogenomic analysis. Trees inferred from proteins conserved across all domains (e.g., ribosomal proteins) that included non-Plantae and/or noncyanobacterial prokaryotes were carefully scrutinized with maximum likelihood (ML) bootstrap analysis of more taxon-rich data sets (i.e., that included nonreference genomes and all significant hits in GenBank and other genome databases) to determine gene origin in the glaucophyte. The final list derived from phylogenomics contains 95 *Cyanophora* trees (genes) with moderate to strong (>70%) bootstrap support for a cyanobacterial gene origin (i.e., 7.7% of the annotated sequences). These proteins were each analyzed in-depth as described above to verify the cyanobacterial ancestry. The protein alignments are available upon request from D.B.

In parallel, we did a BLASTX analysis of the 1,226 *Cyanophora* genes (E value cutoff $< 10^{-10}$) against the nonredundant GenBank database to identify sequences with at least four cyanobacterial homologs among the top five hits. Genes with a significant BLAST hit against a single cyanobacterium were also retained. This approach identified 118 genes of cyanobacterial origin in *Cyanophora*. The combined comprehensive BLAST and phylogenomic analyses returned 132 cyanobacterial genes in *Cyanophora*, corresponding to 10.8% of the 1,226 annotated gene set (Figure 2). False negatives were minimized by using our approach. Because of their short length (< 70 amino acids), we found 37/132 of the proteins through BLAST but not phylogenomics. In contrast 14/132 proteins were found through phylogenomics but were missed by BLAST because they did not satisfy the criteria of 4/5 top hits to cyanobacteria. However, in the latter case, all protein trees with multiple taxa that included Plantae or other algal lineages (chromalveolates, chlorarachniophytes) were sisters to cyanobacteria. To

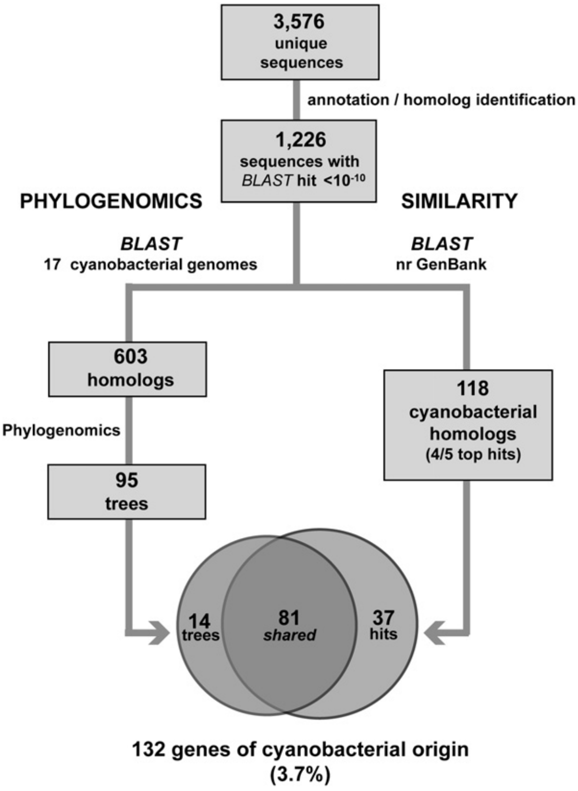

Figure 2. Summary of Phylogenomic and Similarity Approaches to Detect EGT in *Cyanophora paradoxa*

Our strategy for identifying genes of cyanobacterial-EGT origin within 3,576 *Cyanophora* unigenes relies on a dual phylogenomic and similarity strategy to maximize accuracy. Detailed phylogenetic analysis of all candidate genes resulted in a final count of 132 cyanobacterial-derived genes (3.7% of the total number of sequences) in *Cyanophora*, with 81 (*shared*) of these identified through both methods, whereas 14 were found exclusively through phylogenomics (*trees*) and 37 were found exclusively through the BLAST search (*hits*).

address the possibility of gene redundancy due to the existence of gene families, we did a similarity analysis of the 3′-UTR of each EST cluster corresponding to the 132 identified genes. This analysis shows that only eight genes (based on 1%–3% DNA-sequence difference) are potential members of a gene family. When this liberal criterion is used, the average size among the 132 genes is 1.1 genes/family (see Table S2).

We determined the putative cellular location of the 132 proteins of cyanobacterial origin in *Cyanophora* by using Predotar V1.03 [21] and TargetP V1.1 [22]. We identified an N-terminal extension in 13 *Cyanophora* proteins (see Table S2). For other proteins, the most closely related homologs in *Arabidopsis* (or *Oryza sativa* when the *Arabidopsis* gene was not available) were used as proxies to infer the cellular target. This approach showed that 101/132 proteins (76.6%) contain a plastid-targeting signal (identified by both Predotar and TargetP in most cases; see Table S2 for list of probabilities). Additionally, 19 proteins in *Cyanophora* lacked the N terminus but had plastid-genome-encoded homologs in red or green algae (including land plants, e.g., dark-operative protochlorophyllide reductase [DPOR] β subunit) or were present only in cyanobacteria (e.g., porphobilinogen deaminase [PBGD]), and their annotation clearly indicated a plastid function (see Table S2 for details). The grand total of 120 (101 + 19) cyanobacterial-derived, plastid-targeted proteins in *Cyanophora* are, as expected, involved in photosynthetic functions (e.g., chlorophyll synthesis, core photosystem or phycobilisome components, electron transport), plastid metabolism (e.g., oxidative stress, amino acid synthesis, lipid processing), plastid maintenance (e.g., peptide processing, organelle division), and transcription and protein translation (Table S2). In contrast, 12 *Cyanophora* proteins that lack an N-terminal extension have potential nonplastid functions. The predicted targets of the *Arabidopsis* homologs indicate that one of these encodes a hypothetical protein that is likely targeted to the mitochondrion, whereas four proteins with plant homologs have unclear targeting predictions. For one protein, 6-phosphogluconate-dehydrogenase *Arabidopsis* homologs have been reported both as cytosolic and plastid targeted. The remaining six proteins (e.g., glucokinase) are involved in apparent nonplastid functions (see Table S2). These estimates need to be verified by using complete gene sequences from *Cyanophora* as well as a target-prediction program that is trained with data from this species.

We compared the number of genes of cyanobacterial origin in *Cyanophora* with data from other algae and plants. However, the existing analyses of the number and origin of plastid-targeted proteins (of cyanobacterial or other origin) in *Arabidopsis* and in other taxa are difficult to compare directly with each other and to our data because of significant differences in the computational approach and the reference genomes that were included in the analysis. The most up-to-date bioinformatic prediction suggests that there are 4,458 plastid-targeted proteins in *Arabidopsis* [23]. In a prior study, Richly and Leister [24], using BLAST and exhaustive protein targeting analyses, found that out of a total of 2,261 predicted or known plastid-proteins in *Arabidopsis*, 880 have cyanobacterial homologs with roles primarily in metabolism, bioenergetics, and transcription. Moreover, it is estimated that out of 857 plastid-targeted proteins of cyanobacterial ancestry that are shared between *Arabidopsis* and *Oryza*, circa 650 constitute the minimal core set of endosymbiotic proteins required for angiosperm plastid function [24]. The remainder of the plastid proteome is derived from the host, the protomitochondrial genome, or lateral gene transfers. And finally, a recent similarity analysis that was used to identify groups of homologous proteins from a diverse set of genomes (homolog-group method) and to identify the proteins shared between Plantae and nine diverse cyanobacteria (i.e., phylogenetic profiling), suggested that there is a total of 1,192 genes of cyanobacterial origin in *Arabidopsis* and 676 genes in the thermoacidophilic red alga, *Cyanidioschyzon merolae*, respectively [25]. It should be noted that although the genome of *Cyanidioschyzon* is highly reduced (16.5 Mbp, 5,331 predicted genes) as a result of its extreme lifestyle, this alga is a free-living obligate autotroph [26] and therefore likely retains the core set of cyanobacterial genes involved in photosynthetic functions and plastid maintenance. Therefore, although presently unsubstantiated, the available data suggest that *Arabidopsis* contains at least 1,192 nuclear genes of cyanobacterial origin [25], with this number going up to the extrapolated value of 4,500 according to Martin et al. [11].

The *Cyanophora* data provide two key insights with regard to plastid-derived EGT. The first is that given 132/3,576 (3.7%) for the unambiguous cyanobacterial contribution and 12,000–15,000 genes in *Cyanophora*, a minimum of 444–555 genes can be traced back to the plastid endosymbiont. Comparison of the 132 genes in *Cyanophora* with other photosynthetic taxa shows that the vast majority of these sequences are common to *Arabidopsis* (103), *Chlamydomonas* (107), and *Cyanidioschyzon* (109), demonstrating their ancient cyanobacterial provenance in the Plantae (Table S2). If, however, we take the approach of Martin et al. [11] and consider that there is likely to be an equal number of undetected genes in *Cyanophora* of cyanobacterial origin as those identified with a BLAST E value cutoff < 10^{-10}, then by extrapolation 10.8% (132/1,226) of all *Cyanophora* genes (i.e., 1,296–1,620; or circa 1,500) have arisen through EGT. This extrapolation is based on the known tendency of phylogenetic methods to fail as sequence divergence rises within an alignment [11, 27]. Given this bias toward underestimation, it is also very likely that the values of 1,192 and 676 genes [25] represent minimal estimates for cyanobacterial EGT in *Arabidopsis* and *Cyanidioschyzon*, respectively.

Given the estimate based on extrapolation of approximately 1,500 anciently transferred cyanobacterial genes in *Cyanophora*, we still have to account for the large difference with this number in *Arabidopsis* (circa 4,500). This disparity may be explained in two ways. The first is that although Martin et al. [11] had three evolutionarily diverse cyanobacteria (*Nostoc*, *Prochlorococcus*, *Synechocystis*) in their analysis (as well as 11 other Eubacteria and four Archaea), they were limited with respect to the eukaryotic data with only yeast in addition to *Arabidopsis*. This genome data set turned up 1,700 cyanobacterial genes out of a total of 9,368 (i.e., circa 18%) that had significant BLAST hits (E value cutoff < 10^{-10}). If the inclusion of additional eukaryotic (in particular Plantae) genomes were to increase disproportionately the size of the gene set with significant noncyanobacterial BLAST hits in *Arabidopsis*, then the percent cyanobacterial contribution in the genome would diminish in size. For example, an earlier analysis of *Arabidopsis* genes in comparison to all available sequence data showed that 17,833 could be classified according to sequence similarity [28]. Second, whereas the 132 cyanobacterial-derived genes in *Cyanophora* appear to exist primarily in single copies, there is evidence for two rounds of whole-genome duplication during the early evolution of angiosperms [29] as well as rampant gene duplications in *Arabidopsis* [30]. The first genome duplication likely occurred near the Jurassic-Cretaceous boundary (circa 145 million years ago), whereas the second occurred after the monocot-dicot divergence (circa 66–109 million years ago [29]). These events, followed by sequence divergence among surviving gene family members, could significantly inflate the estimate for plastid-derived EGT in *Arabidopsis*. Addressing the potential impact on plastid-derived EGT of more ancient Plantae genome duplications that occurred hundreds of millions of years ago will be a far more challenging issue. If, however, gene duplications did not significantly alter the size of the core set of cyanobacterial genes that was established prior to the divergence of the Plantae lineages (e.g., circa 1.4 billion years ago for the split of red and green algae [1]), then approximately 1,500 plastid-derived genes were present in the common ancestors of the red, green, and glaucophyte algae. Assuming that Martin et al.'s [11] results are correct, then significant increases in the cyanobacterial gene set and their recruitment for nonplastid functions likely occurred sometime during plant evolution, a hypothesis that can be tested by studying earlier-diverging "green" lineages such as multicellular charophyte algae and bryophytes.

Another finding of our study is that the captured plastid endosymbiont apparently provided far fewer genes involved in nonplastid functions in *Cyanophora* than has been postulated (50% [11]) for *Arabidopsis*. We find only 12/132 (9.1%) of the confirmed cyanobacterial genes in this category in *Cyanophora*. This result is critical because understanding ancient EGT addresses more generally the propensity of eukaryotes to incorporate foreign genes into their genomes. EGT is a special case of horizontal gene transfer, with the critical difference that the foreign genes are resident for a protracted period of time in the host cell. If our results are accurate, then the most obvious pool of prokaryotic genes was only minimally tapped for its genetic resources in *Cyanophora*. This again suggests that significant differences exist in different Plantae in the history of diversification, recruitment, and loss of cyanobacterial genes of endosymbiotic origin that are not involved in plastid functions. The amplification of genes of cyanobacterial origin in plants through genome or repeated gene duplications would invariably provide a fertile source of innovation to evolve plastid-independent functions. Such forces likely account (at least partially) for the differences we have observed in the number of non-plastid-targeted proteins of cyanobacterial origin in algae in contrast to their morphologically complex plant sisters.

Plastid establishment provided eukaryotes with oxygenic photosynthesis and a cellular compartment for other key functions such as fatty-acid, isoprenoid, and amino acid biosynthesis. However, the incorporation of cyanobacterial genes into the host metabolism is apparently dissimilar among Plantae lineages and remains to be systematically studied in other members of this group. It is clear that comparison of three distantly related genomes (a unicellular mesophile [*Cyanophora*], an extremophile [*Cyanidioschyzon*], and an angiosperm [*Arabidopsis*]) is not sufficient to understand fully the extent of ancient plastid-derived EGT and its role in the rise of the first photosynthetic eukaryotes. Complete genome sequences from a number of mesophilic, nonpicoeukaryotic, early-diverging Plantae (e.g., bangiophyte-red and prasinophyte-green algae) will be required for this purpose in addition to a complete genome sequence from *Cyanophora* and other glaucophytes.

Experimental Procedures

cDNA Library Construction

Total RNA from a culture of *Cyanophora paradoxa* Pringsheim strain (CCMP329) was extracted by using Trizol (GibcoBRL) and the mRNA purified was by using the Oligotex mRNA Midi Kit (Qiagen). Starter and normalized cDNA libraries were constructed according to Bonaldo et al. [31]. The cDNA clones were sequenced from the 3′ end. Clustering of the 11,176 clones into 3,576 nonredundant sets was performed with the program UIcluster2 [32]. All *C. paradoxa* EST sequences have been deposited in the dbEST database of GenBank.

Phylogenomic Analysis

The EST data were used as input for the phylogenomics approach (for details, see [20, 33]) with the PhyloGenie package of computer programs [34]. PhyloGenie is used for "high-throughput" phylogenetic reconstruction with an automated pipeline in which BLAST searches, extraction of homologous sequences from the BLAST results, generation of protein alignments, phylogenetic-tree reconstruction, and bootstrap support values for individual phylogenies are rapidly calculated. The fast but less robust neighbor-joining method is implemented in PhyloGenie; therefore all candidate trees were subsequently studied with the maximum-likelihood method to verify the results. The local database included 17 completed cyanobacterial genomes or a collection of cyanobacterial, other bacterial, and eukaryotic genomes (see Table S1).

Detailed Phylogenetic Analysis

For all data sets of interest, a single-protein phylogeny was reconstructed under maximum likelihood (ML) by using the PHYML V2.4.3 computer program [35] with the JTT + I + Γ evolutionary model and tree optimization. The α value for the Γ distribution was calculated by using four rate categories. To assess the stability of monophyletic groups in the ML trees, we calculated PHYML bootstrap (100 replicates) support values [36]. The phylogenetic analyses identified 132 genes of cyanobacterial origin in *Cyanophora* (see Table S2).

Analysis of Plastid Targeting in *Cyanophora* Proteins

To assess the probability of plastid targeting for the 132 cyanobacterial-derived proteins in *Cyanophora*, we generated bioinformatics predictions for each protein with the Predotar V1.03 [21] and TargetP V1.1 [22] servers. The results of this analysis using the default values for each program are shown in Table S2.

Supplemental Data

Supplemental Data include two tables and are available with this article online at: http://www.current-biology.com/cgi/content/full/16/23/2320/DC1/.

Acknowledgments

This work was primarily supported by the National Aeronautics and Space Administration and, secondarily, by the National Science Foundation in grants awarded to D.B. (NNG04GM17G, EF 04-31117, respectively). We also acknowledge the constructive criticisms of two anonymous reviewers.

Received: August 23, 2006
Revised: September 19, 2006
Accepted: September 27, 2006
Published: December 4, 2006

References

1. Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G., and Bhattacharya, D. (2004). A molecular timeline for the origin of photosynthetic eukaryotes. Mol. Biol. Evol. *21*, 809–818.
2. Blair, J.E., Shah, P., and Hedges, S.B. (2005). Evolutionary sequence analysis of complete eukaryote genomes. BMC Bioinformatics *6*, 53–63.
3. Hackett, J.D., Yoon, H.S., Butterfield, N.J., Sanderson, M.J., and Bhattacharya, D. Plastid endosymbiosis: Sources and timing of the major events. In Evolution of Aquatic Photoautotrophs, P. Falkowski and A. Knoll, eds. (Academic Press), in press.
4. Bhattacharya, D., and Medlin, L. (1995). The phylogeny of plastids, A review based on comparisons of small subunit ribosomal RNA coding regions. J. Phycol. *31*, 489–498.
5. Palmer, J. (2003). The symbiotic birth and spread of plastids, how many times and whodunit? J. Phycol. *39*, 4–11.
6. Rodriguez-Ezpeleta, N., Brinkmann, H., Burey, S.C., Roure, B., Burger, G., Loffelhardt, W., Bohnert, H.J., Philippe, H., and Lang, B.F. (2005). Monophyly of primary photosynthetic eukaryotes, green plants, red algae, and glaucophytes. Curr. Biol. *5*, 1325–1330.
7. Adl, S.M., Simpson, A.G., Farmer, M.A., Andersen, R.A., Anderson, O.R., Barta, J.R., Bowser, S.S., Brugerolle, G., Fensome, R.A., Fredericq, S., et al. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J. Eukaryot. Microbiol. *52*, 399–451.
8. Martin, W., Rotte, C., Hoffmeister, M., Theissen, U., Gelius-Dietrich, G., Ahr, S., and Henze, K. (2003). Early cell evolution, eukaryotes, anoxia, sulfide, oxygen, fungi first (?), and a tree of genomes revisited. IUBMB Life *55*, 193–204.
9. Martin, W., and Herrmann, R.G. (1998). Gene transfer from organelles to the nucleus, how much, what happens, and why? Plant Physiol. *118*, 9–17.
10. Reumann, S., Inoue, K., and Keegstra, K. (2005). Evolution of the general protein import pathway of plastids. Mol. Membr. Biol. *22*, 73–86.
11. Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002). Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. USA *99*, 12246–12251.
12. Löffelhardt, W., Bohnert, H.J., and Bryant, D.A. (1997). The complete sequence of the Cyanophora paradoxa cyanelle genome. In Origins of Algae and their Plastids, D. Bhattacharya, ed. (Wien, Germany: Springer), pp. 149–162.
13. Guiry, M.D. (2006). AlgaeBase version 4.1. (Galway, Ireland: National University of Ireland) (http://www.algaebase.org).
14. McFadden, G.I. (2001). Primary and secondary endosymbiosis and the origin of plastids. J. Phycol. *37*, 951–959.
15. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO, A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674–3676.
16. Gene Ontology Consortium (2006). The Gene Ontology (GO) project in 2006. Nucleic Acids Res. *34*, D322–D326.
17. Joint Genome Institute, U.S. Department of Energy. (2004) (http://genome.jgi-psf.org/thaps1).
18. Joint Genome Institute, U.S. Department of Energy. (2004) (http://genome.jgi-psf.org/Chlre3).
19. Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., et al. (2004). The genome of the diatom *Thalassiosira pseudonana*, ecology evolution, and metabolism. Science *306*, 79–86.
20. Li, S., Nosenko, T., Hackett, J.D., and Bhattacharya, D. (2006). Phylogenomic analysis identifies red algal genes of endosymbiotic origin in the chromalveolates. Mol. Biol. Evol. *23*, 663–674.
21. Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004). Predotar, A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics *4*, 1581–1590.
22. Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. *300*, 1005–1016.
23. Michigan State University. (2006). Chloroplast 2010. (http://www.plastid.msu.edu/about/gene_list.html).
24. Richly, E., and Leister, D. (2004). An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. Gene *329*, 11–16.
25. Sato, N., Ishikawa, M., Fujiwara, M., and Sonoike, K. (2005). Mass identification of chloroplast proteins of endosymbiotic origin by phylogenetic profiling based on organism-optimized homologous protein groups. Genome Informatics *16*, 56–68.
26. Barbier, G., Oesterhelt, C., Larson, M.D., Halgren, R.G., Wilkerson, C., Garavito, R.M., Benning, C., and Weber, A.P. (2005). Comparative genomics of two closely related unicellular thermo-acidophilic red algae, *Galdieria sulphuraria* and *Cyanidioschyzon merolae*, reveals the molecular basis of the metabolic

flexibility of *Galdieria sulphuraria* and significant differences in carbohydrate metabolism of both algae. Plant Physiol. *137*, 460–474.

27. Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., and Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. *54*, 743–757.
28. Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796–815.
29. De Bodt, S., Maere, S., and Van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. Trends Ecol. Evol. *20*, 591–597.
30. Casneuf, T., De Bodt, S., Raes, J., Maere, S., and Van de Peer, Y. (2006). Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. Genome Biol. *7*, R13.
31. Bonaldo, M.F., Lennon, G., and Soares, M.B. (1996). Normalization and subtraction, two approaches to facilitate gene discovery. Genome Res. *6*, 791–806.
32. Trivedi, N., Bishof, J., Davis, S., Pedretti, K., Scheetz, T.E., Braun, T.A., Roberts, C.A., Robinson, N.L., Sheffield, V.C., Soares, M.B., et al. (2002). Parallel creation of non-redundant gene indices from partial mRNA transcripts. Future Gen. Comp. Sys. *18*, 863–870.
33. Reyes-Prieto, A., Yoon, H.S., and Bhattacharya, D. (2006). Phylogenomics and its growing impact on algal phylogeny and evolution. Algae *21*, 1–10.
34. Frickey, T., and Lupas, A.N. (2004). PhyloGenie, automated phylome generation and analysis. Nucleic Acids Res. *32*, 5231–5238.
35. Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696–704.
36. Felsenstein, J. (1985). Confidence limits on phylogenies, An approach using the bootstrap. Evolution Int. J. Org. Evolution *39*, 783–791.