

# Nucleus-Encoded Periplastid-Targeted EFL in Chlorarachniophytes

Gillian H. Gile and Patrick J. Keeling

Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

Chlorarachniophytes are cercozoan amoeboflagellates that acquired photosynthesis by enslaving a green alga, which has retained a highly reduced nucleus called a nucleomorph. The nucleomorph lacks many genes necessary for its own maintenance and expression, suggesting that some genes have been moved to the host nucleus and their products are now targeted back to the periplastid compartment (PPC), the reduced eukaryotic cytoplasm of the endosymbiont. Protein trafficking in chlorarachniophytes is therefore complex, including nucleus-encoded plastid-targeted proteins, nucleomorph-encoded plastid-targeted proteins, and nucleus-encoded periplastid-targeted proteins. A major gap in our understanding of this system is the PPC-targeted proteins because none have been described in any chlorarachniophytes. Here we describe the first such protein, the GTPase EFL. EFL was characterized from 7 chlorarachniophytes, and 2 distinct types were found. One is related to foraminiferan EFL and lacks an amino-terminal extension. The second, distantly related, type encodes an amino-terminal extension consisting of a signal peptide followed by sequence sharing many characteristics with transit peptides from nucleus-encoded plastid-targeted proteins and which we conclude is most likely PPC targeted. Western blotting with antibodies specific to putative host and PPC-targeted EFL from the chlorarachniophytes *Bigeloviella natans* and *Gymnochlora stellata* is consistent with posttranslational cleavage of the leaders from PPC-targeted proteins. Immunolocalization of both proteins in *B. natans* confirmed the cytosolic location of the leaderless EFL and a distinct localization pattern for the PPC-targeted protein but could not rule out a plastid location (albeit very unlikely). We sought other proteins with a similar leader and identified a eukaryotic translation initiation factor 1 encoding a bipartite extension with the same properties. Transit peptide sequences were characterized from all 3 classes of targeted protein by comparing all examples of each class from expressed sequence tag surveys of *B. natans* and *G. stellata*. No recognizable difference between plastid- and PPC-targeted proteins was observed, but nucleomorph-encoded transit peptides differ, likely reflecting high AT content of nucleomorph genomes. Taken together, the data suggest that the system that directs proteins to the PPC in chlorarachniophytes uses a bipartite targeting sequence, as does the PPC-targeting system that evolved independently in cryptomonads.

## Introduction

One of the most important steps in the transition from endosymbiont to organelle is the establishment of a protein-targeting system. As endosymbionts integrate, many genes are transferred to the host nucleus and those whose products are required in the plastid acquire targeting sequences that are recognized by a specific import apparatus. The targeting system of primary plastids such as those of green algae and plants has been relatively well studied, and most proteins are recognized via an amino-terminal extension known as a transit peptide. Transit peptides tend to share an overall positive charge due to a marked depletion in acidic residues and a modest enrichment in basic residues. Further generalizations can be made for specific subsets of photosynthetic eukaryotes; for example, transit peptides of land plants and to a lesser extent green algae are enriched in serine and threonine, but in general, rules for one lineage may not apply to others.

Since the origin of plastids by primary endosymbiosis, plastids have subsequently moved between eukaryotic lineages by secondary and tertiary endosymbioses. Whereas primary plastids are bound by 2 membranes and located in the host cytosol, secondary and tertiary plastids are bounded by additional membranes and are located within the endomembrane system of the host. As a result, plastid-targeted proteins in secondary algae use a bipartite leader consisting of a signal peptide followed by a transit peptide (McFadden 1999; Patron and Waller 2007). The signal peptide allows proteins to cross the outermost mem-

brane, which is part of the host endomembrane system (and in some taxa is detectably continuous with the endoplasmic reticulum), and the transit peptide is thought to mediate transfer across the 2 innermost membranes, which correspond to the 2 membranes of the primary plastid.

However, most secondary plastids (euglenids and dinoflagellates being the exceptions) have an additional membrane between the outer membrane and the 2 primary plastid membranes, which is thought to be derived from the plasma membrane of the endosymbiotic primary alga. How proteins cross this second membrane is the most poorly understood step of the system. This is partly because there is no obvious leader domain that mediates passage through this membrane and partly because very few proteins are targeted across just the outer and second membranes. Such proteins might allow the requirements for each step of the process to be dissected, but in most algal lineages, few proteins would be expected to function between the 2 pairs of plastid membranes. The 2 exceptions to this are chlorarachniophytes and cryptomonads. In all other secondary algae, the nucleus of the eukaryotic endosymbiont has been completely lost, but in these 2 lineages, relict nuclei called nucleomorphs have been retained in the reduced eukaryotic cytoplasm known as the periplastid compartment (PPC) that lies between the inner and outer pairs of plastid membranes. The majority of nucleomorph genes in both lineages specify housekeeping functions, but in both cases, many genes deemed to be essential for nucleomorph maintenance and expression are missing (Douglas et al. 2001; Gilson et al. 2006). These genes are believed to have been transferred to the host nucleus, from which their products would have to be targeted to the PPC. Thus, a system dominated in most secondary algae by a single plastid-targeting route (host nucleus to plastid) is fragmented into 3 routes in chlorarachniophytes and

Key words: protein targeting, plastids, chlorarachniophytes, transit peptides, *Gymnochlora stellata*.

E-mail: pkeeling@interchange.ubc.ca.

*Mol. Biol. Evol.* 25(9):1967–1977. 2008

doi:10.1093/molbev/msn147

Advance Access publication July 3, 2008

cryptomonads: nucleus-encoded proteins are targeted to 2 distinct compartments (PPC and plastid), and nucleomorph-encoded proteins are targeted to the plastid (fig. 1). Each of these routes has to be specified by targeting information, and the targeting information of PPC-targeted proteins has to be distinguishable from that of plastid-targeted proteins. By crossing only the first and second plastid membranes, the PPC-targeted proteins offer an opportunity to examine how proteins cross the second membrane that is rare or impossible in other lineages, although there is evidence that a few proteins are still targeted to this compartment in diatoms and apicomplexans (Sommer et al. 2007).

In cryptomonads, several such proteins have been identified and clues as to how they are targeted have emerged. These proteins are preceded by an extensive bipartite leader that is intriguingly similar to a plastid-targeting leader: a signal peptide is followed by a sequence that shares many characteristics with cryptomonad transit peptides (Gould, Sommer, Hadfi, et al. 2006). This suggests that the transit peptide, or some variation of it, directs proteins across not only the 2 innermost membranes but also the second membrane in 4-membrane plastids. The transit peptides of cryptomonads, red algae, and other secondary algae with red-derived plastids (possibly all but green algae and plants) tend to have a phenylalanine residue at or near the +1 position (Gould, Sommer, Hadfi, et al. 2006; Patron and Waller 2007). This is thought to derive from the ancient receptor/pore function of the cyanobacterial Omp85, a homolog of the plastid outer membrane translocon protein Toc75, that also requires a phenylalanine residue to bind its substrate outer membrane precursor proteins (Wunder et al. 2007). It has also been demonstrated that a phenylalanine residue is necessary for protein import into glaucophyte cyanelles (Steiner and Löffelhardt 2005), and the lack of it in diatom transit peptides results in targeting to a blob-like structure interpreted as the PPC (Gould, Sommer, Kroth, et al. 2006). It is therefore significant that all the transit peptides of PPC-targeted proteins so far identified in cryptomonads lack this residue, and it has been suggested to play a key role in distinguishing plastid from PPC targeting (Gould, Sommer, Kroth, et al. 2006).

No PPC-targeted proteins have been identified in chlorarachniophytes, and it remains to be seen by what mechanism they cross the second membrane and by what mechanism the cells distinguish between plastid- and PPC-targeted proteins. This is significant because the outermost membrane is crossed using the endomembrane/secretion system of the host and the 2 innermost membranes are crossed using the plastid import system of the primary endosymbiont, so the crossing of the second membrane is the only part of the system that may have evolved completely independently in chlorarachniophytes and cryptomonads. Here we describe transit peptides from EFL, the first putatively PPC-targeted protein to be identified in chlorarachniophytes, and compare them to those of nucleus- and nucleomorph-encoded plastid-targeted proteins from 2 distantly related chlorarachniophytes.

EFL is a GTPase that is related to the translation elongation factor EF-1 $\alpha$  and is thought to have taken over its essential role in translation in several eukaryotic lineages

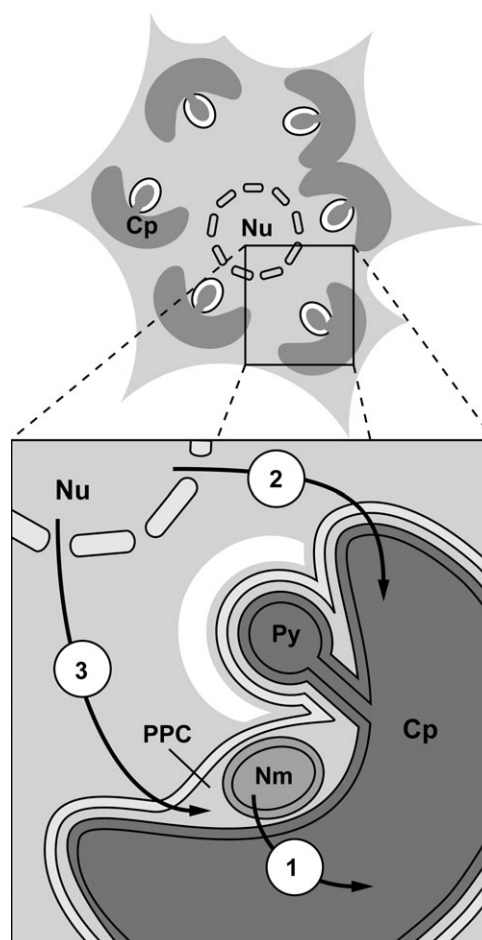


FIG. 1.—Schematic view of plastid targeting in *Gymnochlorella stellata*. Arrows represent routes of protein targeting: 1) nucleomorph-encoded (Nm) proteins are targeted to the plastid (Cp); 2) nucleus-encoded (Nu) proteins are targeted to the plastid; and 3) nucleus-encoded proteins are targeted to the reduced eukaryotic cytosol of the plastid, the PPC.

(Keeling and Inagaki 2004; Gile et al. 2006; James et al. 2006; Ruiz-Trillo et al. 2006; Noble et al. 2007). The chlorarachniophyte *Bigelowiella natans* was previously shown to possess a nucleus-encoded EFL presumed to be of host ancestry (Keeling and Inagaki 2004), and a survey of EFL and EF-1 $\alpha$  in the green algae showed that the ancestor of the chlorarachniophytes' endosymbiont most likely encoded EFL as well (Noble et al. 2007). The *B. natans* nucleomorph genome encodes neither EFL nor EF-1 $\alpha$  (Gilson et al. 2006), which suggests that an endosymbiont-derived EFL has transferred to the host genome and its product is PPC-targeted. Accordingly, we identified 2 evolutionarily distinct clades of EFL in chlorarachniophytes, and we here present evidence that one is a host protein and the other is targeted to the PPC. The putative PPC-targeted proteins include substantial amino-terminal bipartite leaders consisting of a signal peptide and a sequence with similarities to chlorarachniophyte transit peptides. Using these characteristics, we sought other potentially PPC-targeted proteins and identified a eukaryotic translation initiation factor (eIF) with a similar leader that is missing from the *B. natans*

nucleomorph genome. Western blotting of both types of EFL shows that the mature PPC-targeted protein is similar in size to the host-derived proteins, suggesting posttranslational removal of their long leaders. Immunolocalization of both proteins in *B. natans* confirmed that the leaderless EFL is cytosolic and showed a distinct localization pattern for the PPC-targeted protein. However, this pattern could not be distinguished from a plastid localization. Altogether, the evidence suggests that chlorarachniophytes have independently adopted the same overall strategy for PPC targeting as cryptomonads, namely the use of a bipartite targeting peptide similar to plastid-targeting peptides.

In order to characterize these PPC-targeting peptides, we compared them with nucleus- and nucleomorph-encoded plastid-targeted proteins from *B. natans* (Rogers et al. 2004; Gilson et al. 2006) and to corresponding classes of proteins we identified in an expressed sequence tag (EST) survey of the deep-branching chlorarachniophyte, *Gymnochlorella stellata*, thereby including leader sequences from both eukaryotic genomes and across chlorarachniophyte diversity. In both species, plastid- and PPC-targeted proteins encoded in the nucleus share many characteristics, whereas nucleomorph-encoded transit peptides differ, likely reflecting the high AT content of nucleomorph genomes.

## Materials and Methods

### Strains and Culture Conditions

*Chlorarachnion reptans* (strain NEPCC 449) and *Lotharella globosa* (strain NEPCC 811) were obtained from the Canadian Center for the Culture of Microorganisms at the University of British Columbia (CCCM). *Bigeloviella natans* (CCMP 621), *Lotharella amoebiformis* (CCMP 2058), *Lotharella vacuolata* (CCMP 240), *G. stellata* (CCMP 2057), and unidentified chlorarachniophyte strain CCMP 1408 were obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP). All cultures were maintained in f/2-Si or K medium at 22 °C on a 12:12 h light:dark cycle.

### DNA/RNA Extraction, Amplification, and Sequencing

Total RNA was extracted from chlorarachniophyte cell pellets using Trizol reagent (Invitrogen, Carlsbad, CA). Genomic DNA was extracted from *B. natans*, *C. reptans*, *G. stellata*, and *L. vacuolata* using the DNeasy Plant Mini Kit (Qiagen, Mississauga, Ontario, Canada). Full-length chlorarachniophyte EFL cDNA sequences were obtained by the following reactions: 1) 3' rapid amplification of cDNA ends (RACE) using nested degenerate forward primers 5'-GTGCGARATGCAYCAY-3' (outer) and 5'-CCGGGCGAYAAAYGTNGG-3' (inner); 2) reverse transcriptase-polymerase chain reaction (RT-PCR) using gene-specific reverse primers designed from the 3' RACE products and different combinations of nested degenerate forward primers 5'-CTGTCGATCGTCATHTYGGN-3', 5'-TCGTTTCGCGTTCTTNTTYTWYATGGA-3', and 5'-GAGGAGCGGAGCGNGGNGTACNAT-3'; and 3) 5' RACE using specific reverse primers designed from the RT-PCR sequences. The incomplete sequence of *Reti-*

*ulomyxa filosa* EFL was downloaded from the NCBI EST database and finished by polymerase chain reaction (PCR) on genomic DNA using degenerate forward primers. *Thalassiosira weissflogii* EFL was amplified from genomic DNA using degenerate primers designed from *Thalassiosira pseudonana* EFL. The FirstChoice RLM-RACE kit (Ambion, Austin, TX) was used for all 3' RACE and 5' RACE reactions. Superscript III One-Step RT-PCR with Platinum Taq (Invitrogen) was used for all RT-PCR reactions. PCR products were cloned using the TOPO-TA cloning kit (Invitrogen) and sequenced in both directions using the BigDye Terminator v. 3.1 (Applied Biosystems, Foster City, CA).

### Phylogenetic Analyses

New and previously published EFL sequences were translated and aligned using MAFFT (Katoh et al. 2002) and edited in MacClade 4.08 (Maddison DR and Maddison WP 2003) to a final matrix of 50 taxa and 518 unambiguously aligned characters. Phylogenetic trees were inferred using maximum likelihood (ML) and Bayesian methods. ML trees were inferred using PhyML 2.4.4 (Guindon and Gascuel 2003) with input trees generated by BIONJ, the Whelan and Goldman (WAG) model of amino acids substitution, and 4 rate categories approximating a gamma distribution plus a proportion of invariant sites. In all, 1,000 bootstrap replicates were performed with PhyML using the  $\alpha$  parameter and proportion of invariant sites estimated from the original tree. MrBayes 3.0 (Ronquist and Huelsenbeck 2003) was used to perform Bayesian analysis using the WAG substitution model with rates assigned by 4 equally probable categories approximating a gamma distribution. One cold and 3 heated chains were run for 2 million generations, sampling one tree every hundred generations. The first 5,000 sampled trees were discarded as burn-in, and subsequent trees were used to compute the 50% majority-rule consensus tree.

### Sequence Analysis of Targeting Leaders

Eight nucleomorph-encoded and 22 nucleus-encoded genes for plastid products were identified from an ongoing EST survey of *G. stellata*. Genes were identified by similarity to the *B. natans* plastid-targeted proteins and by searching for EST clusters encoding full-length proteins with bipartite targeting sequences at the amino terminus. Several were truncated at the 5' end, and these were completed by 5' RACE as described above. Seventeen nucleomorph-encoded plastid-targeted genes from the complete *B. natans* nucleomorph genome (Gilson et al. 2006) and the 45 nucleus-encoded plastid-targeted genes from a previous EST survey (Rogers et al. 2004) were analyzed in parallel for comparison. Putative transit peptides were analyzed for amino acid content and hydrophobicity. Sliding-window plots of acidic, basic, and hydroxylated amino acids and hydrophobicity profiles were generated as described previously (Rogers et al. 2004) using a window size of 5 residues. For nucleomorph-encoded transit peptides, residues 3–23 were analyzed. Nucleus-encoded transit peptides were aligned at the signal peptide cleavage point predicted by SignalP 3.0



(Bendtsen et al. 2004), and 15 residues upstream and 20 residues downstream were considered for analysis.

Amino acid frequencies of transit peptides were calculated and included as a supplementary table (Supplementary Material online). Mature plastid-targeted and cytosolic protein data sets were assembled, and their amino acid frequencies were also computed in order to provide a point of comparison to the transit peptides. Nucleomorph-encoded mature protein data sets consisted of residues 100 to the end of the shortest of the plastid-targeted proteins, thereby including 83 residues from each of 8 proteins for *G. stellata* and 84 residues from 17 proteins for *B. natans*. The nucleus-encoded mature protein data sets included residues from 100 to the end of all plastid-targeted proteins. Twenty-eight nucleus-encoded cytosolic proteins were also identified in the *G. stellata* ESTs and compared against the 38 nucleus-encoded cytosolic proteins from the previous *B. natans* EST survey (Rogers et al. 2004). In addition, 16 nucleomorph-encoded nontargeted proteins from *G. stellata* were compared with their homologs in the *B. natans* nucleomorph genome. New sequences from this study were deposited in GenBank under accession numbers EU810236–EU810337.

#### Immunoblotting and Localization

Polyclonal antibodies were raised against a mixture of 2 synthetic peptide sequences unique to the putatively PPC-targeted EFL but conserved among diverse chlorarachniophytes, CDQAKYKEERYNEILK and KETGGKKVEDPKMLK (BioSynthesis Inc., Lewisville, TX) and against 2 synthetic peptides from the *B. natans* cytosolic EFL, CIVGVNKMDEKSVKYD and GKITDCKNNPVKTVS (AbCam, Cambridge, United Kingdom). The *B. natans* cytosolic EFL antibodies were affinity purified before use. Cells were harvested from cultures of *G. stellata*, *B. natans*, and *L. vacuolata*, pelleted by centrifugation, resuspended in 0.5 ml lysis buffer (50 mM Tris pH 7.5, 200 mM sorbitol, and 1 mM ethylenediaminetetraacetic acid) with 5  $\mu$ l protease inhibitor cocktail (Sigma-Aldrich, St Louis, MO) and 10  $\mu$ l phenylmethylsulfonyl fluoride in isopropanol (20 mg/ml), and repeatedly shock frozen in liquid N<sub>2</sub> to release the proteins. Cell lysates were added to sample buffer and boiled for 15 min before separation by Sodium dodecyl sulfate–polyacrylamide gel electrophoresis on a 10% Tris–glycine gel. Separated proteins were transferred to Hybond-P PVDF transfer membrane (Amersham, Buckinghamshire, UK) for 70 min at 100 V. Membranes were blocked and incubated with primary antiserum at 1:1,000 and then a peroxidase-conjugated goat anti-rabbit IgG antibody (BioRad, Hercules, CA) at 1:3,000 dilution. Blots were visualized using the enhanced chemiluminescence detection system (Amersham).

For localization experiments, *B. natans* cells were fixed in cold 4% paraformaldehyde, settled on cover glass slides coated with Histogrip (Invitrogen), and blocked with 5% bovine serum albumin (BSA) at room temperature, shaking, for 1 h. Primary antibodies were applied at a concentration of 1:200 in 5% BSA for 1 h, shaking, at room temperature. Slides were washed, incubated with Alexa Fluor 488 goat anti-rabbit IgG (Molecular Probes, Carlsbad, CA) at a concentration of 1:1,000, washed again,

mounted using ProLong Gold antifade reagent with 4',6-diamidino-2-phenylindole (DAPI) (Invitrogen) according to the manufacturer's instructions, and observed on an Axioplan 2 compound microscope (Zeiss, Göttingen, Germany) with an AttoArc 2 100 W Mercury lamp (Zeiss) for fluorescence visualization.

## Results

### Chlorarachniophyte Host Nuclei Encode 2 Distinct Clades of EFL

The 19 EFL sequences characterized from 7 chlorarachniophyte strains group into 2 distantly related clades. One clade (referred to as “cytosolic” in fig. 2) is made up of relatively divergent sequences that share several distinguishing indels and includes the EFL sequences found in both *B. natans* and *G. stellata* EST libraries. Some species were found to encode 2 or 3 paralogs of this gene, some indicating recent duplications and others suggesting more ancient duplications. The clade is well supported overall (100%) and is related with weak support to the foraminiferan *R. filosa*, with which it shares 2 otherwise unique insertions. Genomic copies of cytosolic EFL were sequenced from *B. natans* and *G. stellata* and were found to completely lack introns. Complete 5' sequences were acquired by 5' RACE from *B. natans*, *L. globosa*, *L. amoebiformis*, and the *L. vacuolata* sequences A and B, and none encodes an extension longer than a few amino acids. Taken together, the evidence points to this gene being derived from the cercozoan host lineage and its product functioning in the host cytoplasm.

The second clade (referred to as “PPC-targeted” in fig. 2) is also well supported (100%) and made up of slightly less divergent EFL sequences represented by a single sequence in each strain. This clade is not demonstrably related to green algal EFL, but the backbone of the tree is completely unresolved, and the green algae themselves do not group into a single clade. The possibility that the gene has a foreign origin cannot be ruled out, as approximately 20% of *B. natans* plastid-targeted proteins are thought to have been acquired by lateral gene transfer (Archibald et al. 2003). However, because EFL is relatively rare among eukaryotes compared with EF-1 $\alpha$  and because the chlorarachniophyte plastid is descended from a group where EFL is found in nearly all major subgroups (Noble et al. 2007), the simplest explanation for the origin of this second gene is that it is derived from the endosymbiont. None of the characterized copies of this gene appears to be encoded in the nucleomorph, however. This is supported by 3 lines of evidence. First, the complete nucleomorph genome of *B. natans* lacks EFL (Gilson et al. 2006). Second, the nucleomorph genome is >65% AT in the single-copy regions (Gilson et al. 2006), but the PPC-targeted EFL sequences are 46.2% AT in *B. natans*, 49.8% in *G. stellata*, and 48.2% in *L. vacuolata*, consistent with the host nuclear genome. Third, and most compelling, nucleomorph introns are well studied and conspicuous: there are over 800 of them in the *B. natans* nucleomorph, and all are between 18 and 21 bp in length (Gilson et al. 2006). Over 100 from *G. stellata* have been characterized and are similarly reduced (Slamovits CH, unpublished data). We amplified and sequenced

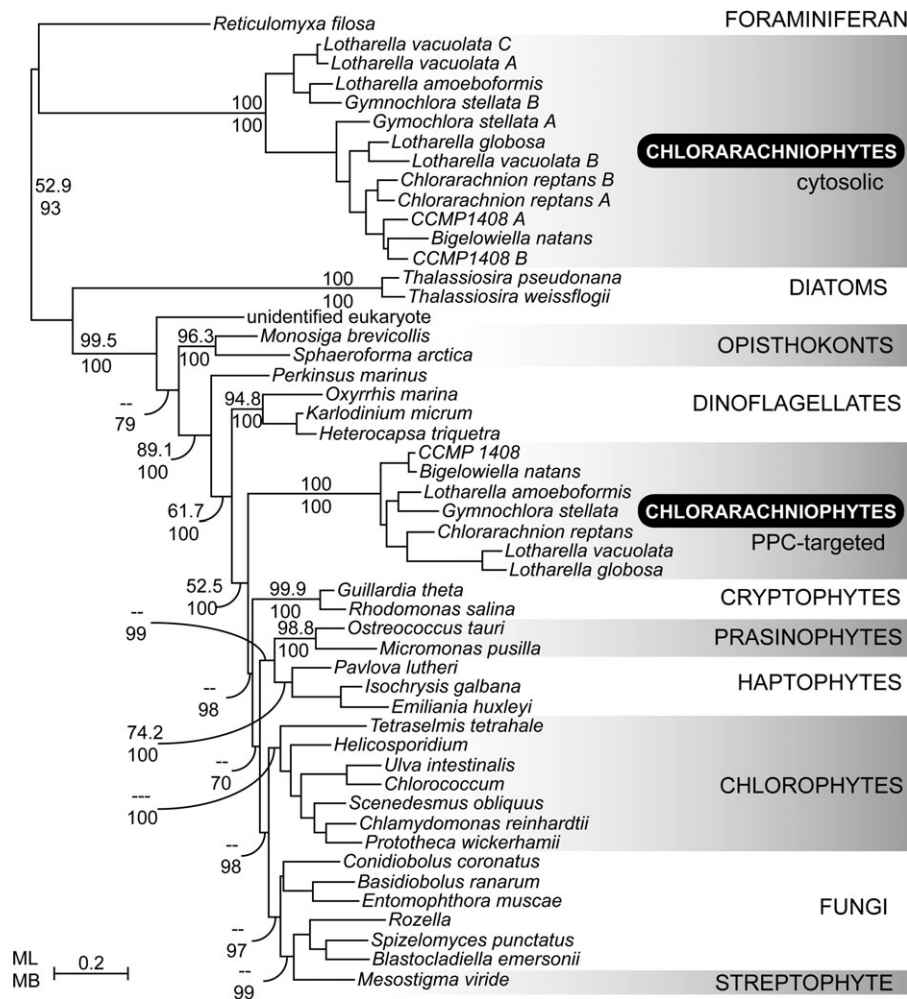


FIG. 2.—Protein ML tree of 50 EFL sequences using 518 sites. Major groups are named to the right, and the inferred functional location of chlorarachniophyte proteins is indicated. Support for nodes greater than 50% is given from ML bootstrap values (above) and Bayesian posterior probability values (below).

portions of the genomic copy of PPC-targeted EFL and found 5 introns in *L. vacuolata* and 3 introns each in *C. reptans* and *B. natans* ranging from 47 to 176 bp in length.

#### Evidence for PPC Targeting of EFL and Characteristics of Transit Peptides

The putative PPC-targeted version of EFL is encoded in the host nucleus, so if it functions in the endosymbiont, it must be targeted. To examine these genes for evidence of targeting peptides, we characterized the 5' end of PPC-targeted genes from *B. natans*, *G. stellata*, and *L. vacuolata* by 5' RACE. In all 3 cases, a bipartite targeting sequence was found, consisting of a signal peptide followed by a transit peptide-like sequence. Hydrophobic signal peptides of 39 amino acids in *B. natans* and 40 in *G. stellata* and *L. vacuolata* were predicted with high support (fig. 3A and B and supplementary fig. S1, Supplementary Material online), as is typical for chlorarachniophyte plastid-targeted proteins (Rogers et al. 2004). The predicted signal cleavage sites are ARQ for *B. natans*, ALA for *G. stellata*, and SFA for *L. vacuolata*, which conform to the expectations for signal

cleavage sites of eukaryotic secreted proteins. Sliding-window plots of the PPC-targeting sequences show extreme levels of hydrophobicity in the signal peptide relative to the plots of plastid-targeting sequences (fig. 3A and B vs. C and D), but this is mainly due to the dampening effect that averaging has on the larger data set of plastid-targeting peptides.

Between the predicted signal cleavage site and the start of sequence conservation with mature EFL proteins is between 59 (in *B. natans*) and 81 (in *L. vacuolata*) amino acids of sequence. With only one putative PPC-targeted protein from each species, characteristics of this class of targeting sequence within each species cannot be generalized, but comparing the *B. natans* and *G. stellata* sequences to transit peptides from their plastid-targeted proteins is informative. A collection of nucleus-encoded plastid-targeted proteins from *B. natans* has been analyzed previously (Rogers et al. 2004), so we developed a comparable set of proteins from *G. stellata*. Twenty-two nucleus-encoded plastid-targeted proteins were identified from a *G. stellata* EST library and, where truncated, the complete sequence of their leaders acquired by 5' RACE. Twenty-eight

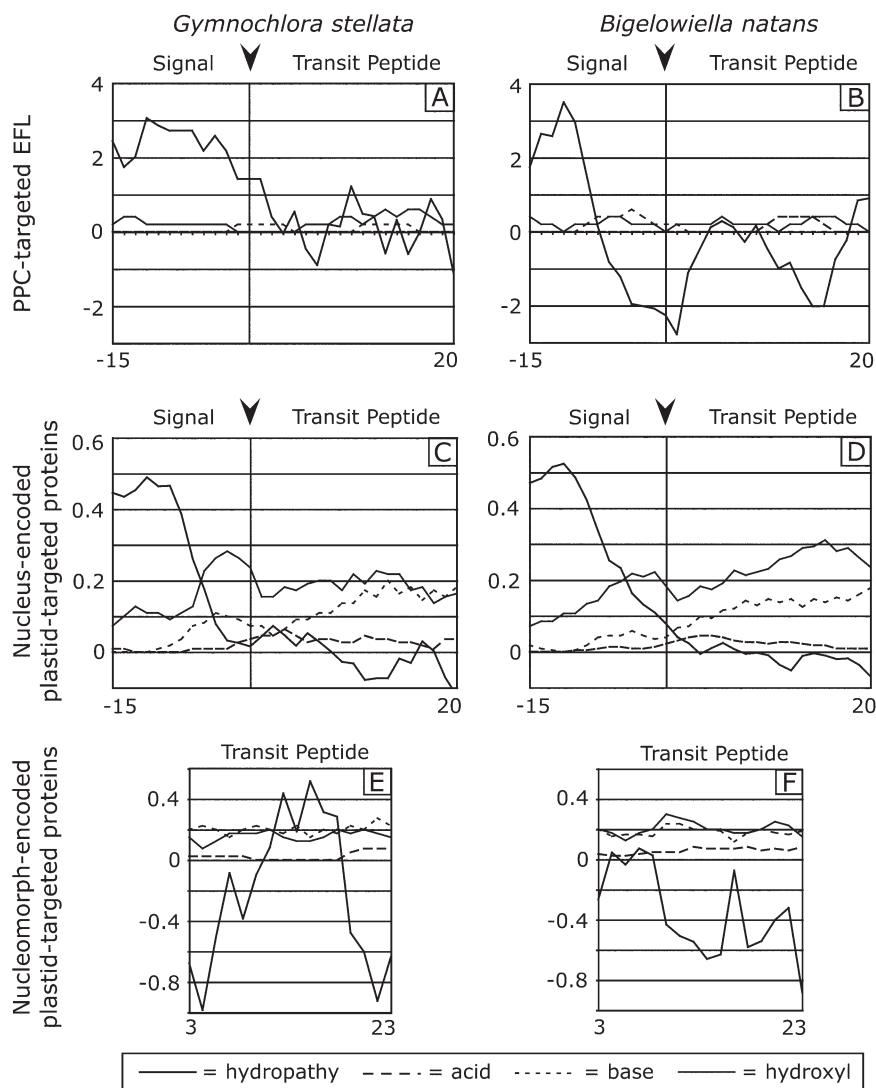


FIG. 3.—Sliding-window plots of signal and transit peptide characteristics of PPC-targeted EFL (A and B), nucleus-encoded plastid-targeted proteins (C and D), and nucleomorph-encoded plastid-targeted proteins (E and F). Hydrophathy profiles are computed by averaging the total of the Kyte–Doolittle hydrophathy scores of the residues in the window over the total window size (here 5 residues). Acid, base, and hydroxyl plots represent the number of residues with that property in the window divided by the size of the window (also 5 residues). Scores are averaged over all transit peptides in that class. Proteins from each class are aligned at their predicted signal cleavage sites (arrowheads).

*G. stellata* cytosolic proteins were also identified and completely sequenced to provide a baseline of amino acid composition of nontargeted sequences. Transit peptide lengths could not be determined unambiguously in all cases, so the first 20 amino acids following the predicted signal cleavage site were analyzed for chemical characteristics. The transit peptides from *G. stellata* and *B. natans* plastid-targeted proteins have remarkably similar amino acid compositions, and most significantly they share characteristics with the transit peptides of the PPC-targeted EFL (fig. 4A–D). In plastid- and PPC-targeted proteins in both species, this region is enriched in serine and arginine relative to both the cytosolic and mature plastid-targeted proteins, although *G. stellata* is further enriched in alanine and proline. They are both depleted in acidic residues as is expected of transit peptides in general (fig. 3C and D), and both are more severely depleted in lysine than either aspartic or glutamic acid. These characteristics are reminiscent of plant transit peptides, which

also tend to be enriched in serine and arginine and depleted in acidic residues. Overall, the N-terminal extensions of the PPC-targeted EFL proteins share all known characteristics of the chlorarachniophyte nucleus-encoded plastid-targeting leaders.

#### Western Blot Results Are Consistent with Post-Import Cleavage of PPC-Targeting Peptides

The predicted leader sequences in PPC-targeted EFL genes are substantial in size (over 100 amino acids), so we analyzed the size of mature PPC-targeted EFL by immunoblotting to determine whether there is any comparable reduction in the apparent size of the mature protein compared with the predicted size of the full-length gene product. Western blots of proteins from *B. natans*, *G. stellata*, and *L. vacuolata* with antibodies raised against peptide sequences specific to the PPC-targeted proteins were compared with blots of

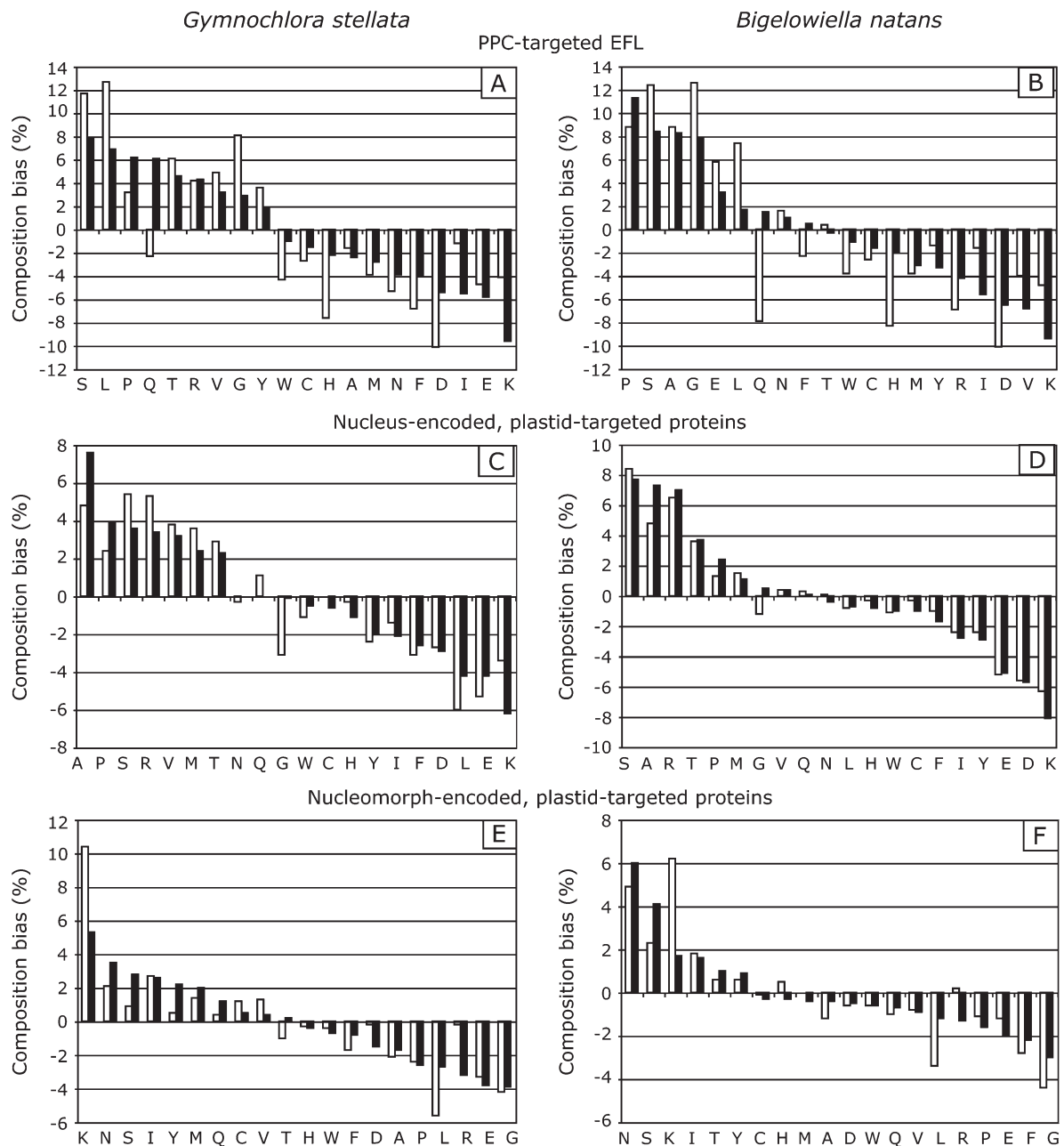


FIG. 4.—Relative amino acid composition of transit peptides of PPC-targeted EFL (A and B), nucleus-encoded plastid-targeted proteins (C and D), and nucleomorph-encoded plastid-targeted proteins (E and F). Bars indicate the difference between amino acid frequencies (%) in transit peptides versus mature targeted proteins (outlined bars) and transit peptides versus cytosolic proteins (solid bars).

*B. natans* proteins reacted to antibodies raised against the *B. natans* cytosolic EFL. In all 3 species, the major PPC-targeted EFL band is significantly smaller than the predicted size of the full-length protein and similar in size to the *B. natans* cytosolic EFL (fig. 5). This is the size we would expect for the mature targeted EFL if the targeting sequence is cleaved. This method lacks the resolution to determine the exact length of the putatively cleaved sequence (assuming the major band is our protein of interest), but size estimates of PPC-targeted proteins suggest that cleavage would take place at or near the stretch of 2–3 glycine residues followed by a stretch of acidic and small neutral residues that immediately precedes

the start of sequence homology to other EFL sequences (supplementary fig. S2, Supplementary Material online). The lengths of the entire bipartite leader sequences would thus be approximately 90 amino acids in *B. natans*, 100 in *G. stellata*, and 110 in *L. vacuolata*.

#### Cytosolic and Targeted EFL Have Distinct Localization Patterns

The cellular location of both EFL proteins was examined in *B. natans* using immunofluorescence and revealed 2 distinct localization patterns (fig. 6). The cytosolic EFL



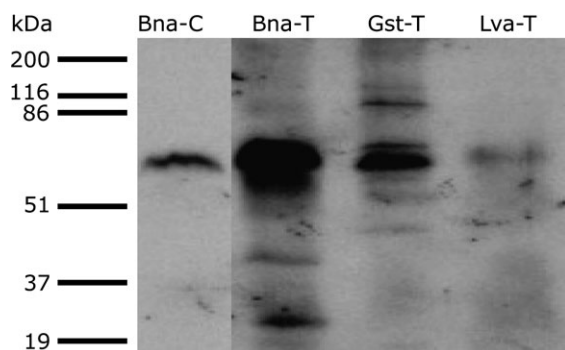


FIG. 5.—Western blots of proteins extracted from *Bigeloviella natans* (Bna), *Gymnochlorella stellata* (Gst), and *Lotharella vacuolata* (Lva) probed with antibodies raised to the *B. natans* cytosolic EFL (C) or epitopes common to the PPC-targeted clade of EFL (T).

protein appears to be distributed throughout the cytoplasm, as expected. The PPC-targeted EFL has a distinct localization pattern, colocalizing with plastid autofluorescence. From transmission electron micrographs, the PPC is known to be a thin layer that completely surrounds the plastid (Gilson and McFadden 1999; Møestrup and Sengco 2001). However, we were unable to detect a staining pattern that clearly surrounded the plastid without also occurring within it, even using confocal microscopy (data not shown). Moreover, the targeted EFL does not appear to colocalize with the nucleomorphs, which are faintly visible by DAPI staining and which would be expected to be surrounded by the largest visible volume of PPC. Assuming the antibody is specific for PPC-targeted EFL, and that the plastid signal is not due to some other artifact of fixation or binding, this observation can be interpreted in a number of ways. First, the PPC-targeted EFL might actually function in the chloroplast. This would be quite remarkable because EFL is so far strictly found in eukaryotic cytoplasm and the *tufA* gene in the chloroplast genome is known to be transcriptionally active by its representation in the *B. natans* EST library. Alternatively, the targeting system may not be sufficiently differentiated to discriminate perfectly between plastid- and PPC-targeted proteins so that PPC proteins are present in both compartments, and the PPC signal is obscured by the plastid signal. Both these scenarios are consistent with the overall resemblance of the targeted EFL leader to plastid-targeting peptides. The other major possibility is that the observed localization pattern does not reflect the actual location of this protein. This could be due to nonspecific binding of our polyclonal antibodies to chloroplast proteins (although BLAST similarity searches of our EFL epitopes against the nucleomorph and chloroplast genomes of *B. natans* fail to return any matches), or it could be an artifact of the localization procedure such as degradation of the chloroplast membranes before fixation. A resolution between these possibilities should be available soon. The *B. natans* genome is currently being sequenced, and analysis of more putatively PPC-targeted proteins from that genome should clarify whether or not there is a distinct class of targeting peptides specific for PPC targeting and if the PPC-targeted EFL leader matches the characteristics of that class. In addition, a transformation system for *L. amoebiformis* has

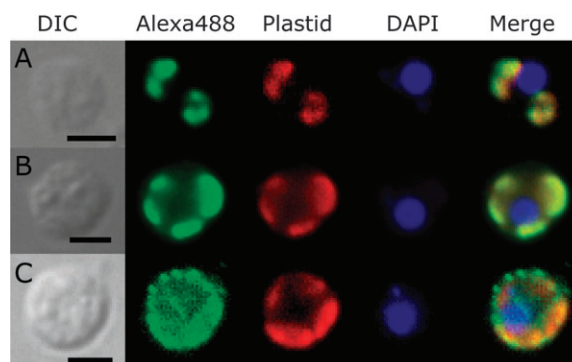


FIG. 6.—Immunolocalization of targeted (A and B) and cytosolic (C) EFL proteins in *Bigeloviella natans*. Scale bar represents 2 μm.

been published recently (Hirakawa et al. 2008), and when putatively PPC-targeted sequences are available from this species or when a transformation protocol is available for *B. natans* or *G. stellata*, green fluorescent protein fusions can be used to determine the location of proteins much more clearly.

#### PPC-targeted eIF1

If the leader on the PPC-targeted EFL does represent the characteristics of PPC-targeted peptides as a whole, then other putatively PPC-targeted proteins might be identifiable based on their possession of similar leaders. Accordingly, an eIF1 was identified in the *G. stellata* EST library by virtue of a similar N-terminal extension to that found on the targeted clade of EFL. The leader consists of a 31-residue signal peptide and a short stretch of amino acids with similar characteristics to chlorarachniophyte transit peptides before the start of the eIF1 domain (supplementary fig. S3, Supplementary Material online). Although the N-terminus of eIF1 is variable and some sequences have N-terminal extensions longer than that found on *G. stellata*, this is the only case in which SignalP strongly predicts a signal peptide. This protein is not likely encoded in the *G. stellata* nucleomorph because its AT content is only 57.7% and because it is missing from the complete *B. natans* nucleomorph genome. Like the PPC-targeted EFL leaders, eIF1 has an acidic stretch near its C-terminus, a characteristic that is uncommon in transit peptides. Although this may represent a recognizable trait by which the cell differentiates between PPC and plastid protein traffic, this possibility will need to be validated experimentally.

#### Characteristics of *G. stellata* and *B. natans* Nucleomorph-Encoded Transit Peptides

The *B. natans* nucleomorph genome encodes 17 annotated genes for plastid-targeted proteins. We identified 8 of these as full-length cDNAs in the *G. stellata* EST survey. Based on the start of sequence similarity between the nucleomorph proteins and their green algal homologs, transit peptides in both species ranged from 23 to 70 amino acids, but cleavage sites could not be unambiguously assigned. To characterize these leaders and determine how they differ



from nucleus-encoded transit peptides, we assembled an additional set of 16 nucleomorph-encoded cytosolic proteins from the *G. stellata* ESTs and their counterparts from the *B. natans* nucleomorph genome.

In general, nucleomorph-encoded transit peptides are enriched in basic residues, especially lysine at 17%, and depleted in acidic residues, especially glutamic acid (fig. 3E and F), relative to the mature proteins. Because the chlorarachniophyte plastid is descended from a green alga, we might expect to find enriched levels of serine and threonine, but they are only somewhat enriched relative to cytosolic proteins and not at all enriched relative to mature plastid-targeted proteins. The clearest trend in these transit peptides is the enrichment in asparagine and lysine (fig. 4E and F and supplementary table T1, Supplementary Material online). This is reminiscent of apicomplexan and cryptomonad nucleomorph transit peptides, where the bias is driven by the high AT content of their genomes favoring amino acids with AT-rich codons (Ralph et al. 2004) and contrasts sharply with the nucleus-encoded transit peptides in which lysine is the most severely depleted residue. Leucine and glycine are the most depleted in chlorarachniophyte nucleomorph transit peptides, even more than the acidic residues.

## Discussion

### Origin and Evolution of Chlorarachniophyte EFL Genes

We have shown that the host nuclear genomes of a diverse sample of chlorarachniophytes encode 2 phylogenetically distinct EFL genes. One is weakly related to foraminiferan relatives of chlorarachniophytes, encodes no targeting information, and is localized to the cytosol in immunolocalization experiments. We conclude that the product of this gene functions in the host cytoplasm. The other gene is not demonstrably related to green algal EFL, but we conclude that its product nevertheless most likely functions in the endosymbiont cytosol for a number of reasons. First, the translation function of EF-1 $\alpha$ /EFL is essential and the endosymbiont likely encoded EFL when it was engulfed, but the relict nucleomorph genome no longer encodes either gene. Second, the PPC-targeted EFL encodes a leader with all the characteristics expected of a plastid-targeting leader in chlorarachniophytes, and immunoblotting indicates that the leader is processed as expected for a targeting peptide. Immunofluorescence localization is consistent with a PPC and/or a plastid location of this protein, but because EFL is restricted to eukaryotes and its presumed function in the plastid is fulfilled by the plastid-encoded *tufA* (Rogers et al. 2007), the only logical compartment in which to assign the second EFL is the endosymbiont cytosol (although whether it also exists in the plastid in *B. natans* in a functional capacity needs to be clarified). The simplest explanation for this is that both host and endosymbiont used EFL at the time they were united, and the endosymbiont gene moved to the host nucleus from which its product is posttranslationally targeted back to the compartment in which it has always functioned. This makes chlorarachniophytes an interesting case as they are a union of 2 cells with the relatively rare EFL protein,

unlike cryptomonads where the host uses EFL but the endosymbiont uses a nucleus-encoded, PPC-targeted EF-1 $\alpha$  (Gould, Sommer, Kroth, et al. 2006).

### Parallel Evolution of PPC Targeting in Chlorarachniophytes and Cryptomonads

The first putative nucleus-encoded PPC-targeted proteins have recently been described in cryptomonads, including EF-1 $\alpha$  (Wastl and Maier 2000; Gould, Sommer, Kroth, et al. 2006; Sommer et al. 2007). Interestingly, a comparison of PPC- and plastid-targeting peptides came to the same conclusion as we reach here: the leaders are composed of signal peptides and transit peptide-like sequences (Gould, Sommer, Kroth, et al. 2006). Because cryptomonads and chlorarachniophytes are very distantly related and their plastids were acquired by independent endosymbiotic events from 2 different primary algal groups, any similarity in their PPC-targeting systems must have evolved in parallel. This is significant because other major steps in the targeting pathway were assembled from existing machinery of the host (signal peptides and endomembrane targeting) or the endosymbiont (transit peptides and the translocons of the inner and outer chloroplast membranes systems, although note that so far only one putative TOC component has been identified in chlorarachniophytes [Gilson et al. 2006] and none in cryptomonads). PPC targeting and the crossing of the second membranes are the only steps that could significantly differ between these independently evolved systems, and yet, it now appears that both groups have arrived at the same solution, namely some modification of the transit peptide. This is more remarkable given the different plastid membrane topology in these 2 groups. The outermost membranes of chlorarachniophyte plastids are smooth and not continuous with the endoplasmic reticulum, whereas cryptomonad plastids reside within the rough endoplasmic reticulum. Thus, the strategy for targeting between the entry to the endomembrane and the entry to the plastid itself could be quite different. It has been hypothesized that the second membrane houses a modified TOC complex that recognizes some transit peptides and not others (Cavalier-Smith 1999), and such a model is entirely consistent with our data for chlorarachniophytes. The only difference at present is that we have not determined the mechanism by which the chlorarachniophyte cell distinguishes between PPC- and plastid-targeting transit peptides. The phenylalanine identified as critical in cryptomonads (Gould, Sommer, Kroth, et al. 2006) was not used by the ancestor of green algae (Patron and Waller 2007), so if this is a key to PPC targeting in cryptomonads, then chlorarachniophytes must have adopted a different key (and perhaps a less stringent key if the distinction between plastid and PPC targeting is relaxed as our initial immunolocalization data suggest). Apicomplexan parasites may provide a useful comparison for understanding the chlorarachniophyte plastid targeting because both groups have 4-membrane-bound plastids with a smooth outer membrane that is not continuous with the endoplasmic reticulum. Although these groups are unrelated, similarities in plastid targeting between *Euglena gracilis* and dinoflagellates show that

plastid membrane topology can influence targeting despite lack of relatedness (Patron et al. 2005, Durnford and Gray 2006). The EFL- and eIF1-targeting sequences share an acidic stretch at their C-termini, but whether this is a general trend and whether it has any functional significance remains to be seen. Targeted chlorarachniophyte EFL sequences also have a hydrophilic sequence of alternating stretches of lysine and aspartic acid residues at their C-termini that is lacking in all other EFL proteins including cytosolic chlorarachniophyte EFL. If this extension is involved in targeting, it would represent a novel mechanism for targeting to a plastid, although peroxisomal proteins are targeted via an uncleaved C-terminal extension. However, this feature is lacking in the putatively PPC-targeted eIF1. The use of bipartite targeting sequences in both chlorarachniophytes and cryptomonads would suggest potential restrictions on the range of possible solutions to the problem of PPC targeting, and in both cases, it was solved by modifying an existing system.

### Three Classes of Transit Peptides in Chlorarachniophytes

Although the nucleus-encoded plastid and PPC-targeting peptides in chlorarachniophytes are unexpectedly similar to one another, the transit peptides of nucleomorph-encoded proteins are remarkably different, despite the fact that these proteins are destined for the same compartment and presumably recognized by the same import complexes. It is possible that the requirements of crossing the second membrane impose a functional constraint on nucleus-encoded transit peptides, which causes them to differ from those of nucleomorph-encoded proteins. More likely, however, this reflects the flexibility of the import system and the nature of the genomes where they are encoded. Cryptomonad nucleomorph-encoded transit peptides are similarly biased toward lysine and asparagine (Ralph et al. 2004), the genomes are similarly AT rich (Douglas et al. 2001; Gilson et al. 2006), and in both cases, the nucleomorph-encoded proteins have lost ancestral characteristics retained by their nucleus-encoded counterparts (elevated serine levels in chlorarachniophytes and the phenylalanine near the N-terminus in cryptomonads). Chlorarachniophyte and cryptomonad nucleomorph-encoded transit peptides are remarkably similar due to a parallel AT bias in the nucleomorph genomes. We have shown that these 2 lineages also have the shared characteristic of using a variation on the bipartite plastid-targeting peptides to target proteins to the PPC.

### Supplementary Material

Supplementary table T1 and figures S1–S3 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

### Acknowledgments

This work was supported by a grant (227301) from the Natural Sciences and Engineering Research Council of Canada (NSERC), and *G. stellata* EST sequencing was sup-

ported by a grant to the Centre for Microbial Diversity and Evolution from the Tula Foundation. We thank C. H. Slamovits for use of the *G. stellata* ESTs. G.H.G. is supported by an NSERC postgraduate doctoral scholarship. P.J.K. is a Fellow of the Canadian Institute for Advanced Research and a Senior Investigator of the Michael Smith Foundation for Health Research.

### Literature Cited

- Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ. 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigeloviella natans*. *Proc Natl Acad Sci USA*. 100:7678–7683.
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: signalP 3.0. *J Mol Biol*. 340:783–795.
- Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol*. 46:347–366.
- Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001. The highly reduced genome of an enslaved algal nucleus. *Nature*. 410:1091–1096.
- Durnford DG, Gray MW. 2006. Analysis of *Euglena gracilis* plastid-targeted proteins reveals different classes of transit sequences. *Eukaryot Cell*. 5:2079–2091.
- Gile GH, Patron NJ, Keeling PJ. 2006. EFL GTPase in cryptomonads and the distribution of EFL and EF-1alpha in chromalveolates. *Protist*. 157:435–444.
- Gilson PR, McFadden GI. 1999. Molecular, morphological and phylogenetic characterization of six chlorarachniophyte strains. *Phycol Res*. 47:7–19.
- Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc Natl Acad Sci USA*. 103:9566–9571.
- Gould SB, Sommer MS, Hadfi K, Zauner S, Kroth PG, Maier UG. 2006. Protein targeting into the complex plastid of cryptophytes. *J Mol Evol*. 62:674–681.
- Gould SB, Sommer MS, Kroth PG, Gile GH, Keeling PJ, Maier UG. 2006. Nucleus-to-nucleus gene transfer and protein retargeting into a remnant cytoplasm of cryptophytes and diatoms. *Mol Biol Evol*. 23:2413–2422.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.
- Hirakawa Y, Kofuji R, Ishida K. 2008. Transient transformation of a chlorarachniophyte alga, *Lotharella amoebiformis* (Chlorarachniophyceae), with *uidA* and *egfp* reporter genes. *J Phycol*. 44:814–820.
- James TY, Kauff F, Schoch CL, et al. (70 co-authors). 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. *Nature*. 443:818–822.
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 30:3059–3066.
- Keeling PJ, Inagaki Y. 2004. A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. *Proc Natl Acad Sci USA*. 101:15380–15385.

- Maddison DR, Maddison WP. 2003. MacClade 4: analysis of phylogeny and character evolution 4.08.
- McFadden GI. 1999. Plastids and protein targeting. *J Eukaryot Microbiol.* 46:339–346.
- Møestrup O, Sengco M. 2001. Ultrastructural studies on *Bigelowiella natans*, gen. et sp. nov., a chlorarachniophyte flagellate. *J Phycol.* 37:624–646.
- Noble GP, Rogers MB, Keeling PJ. 2007. Complex distribution of EFL and EF-1 $\alpha$  proteins in the green algal lineage. *BMC Evol Biol.* 7:82.
- Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: a global analysis of plastid targeting signals. *Bioessays.* 29:1048–1058.
- Patron NJ, Waller RF, Archibald JM, Keeling PJ. 2005. Complex protein targeting to dinoflagellate plastids. *J Mol Biol.* 348:1015–1024.
- Ralph SA, Foth BJ, Hall N, McFadden GI. 2004. Evolutionary pressures on apicoplast transit peptides. *Mol Biol Evol.* 21:2183–2194.
- Rogers MB, Archibald JM, Field MA, Li C, Striepen B, Keeling PJ. 2004. Plastid-targeting peptides from the chlorarachniophyte *Bigelowiella natans*. *J Eukaryot Microbiol.* 51:529–535.
- Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. 2007. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol Biol Evol.* 24:54–62.
- Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* 19:1572–1574.
- Ruiz-Trillo I, Lane CE, Archibald JM, Roger AJ. 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaerofarma arctica*. *J Eukaryot Microbiol.* 53:379–384.
- Sommer MS, Gould SB, Lehmann P, Gruber A, Przyborski JM, Maier UG. 2007. Der1-mediated preprotein import into the periplastid compartment of chromalveolates? *Mol Biol Evol.* 24:918–928.
- Steiner JM, Löffelhardt W. 2005. Protein translocation into and within cyanelles. *Mol Membr Biol.* 22:123–132.
- Wastl J, Maier UG. 2000. Transport of proteins into cryptomonads complex plastids. *J Biol Chem.* 275:23194–23198.
- Wunder T, Martin R, Löffelhardt W, Schleiff E, Steiner JM. 2007. The invariant phenylalanine of precursor proteins discloses the importance of Omp85 for protein translocation into cyanelles. *BMC Evol Biol.* 7:236.

Martin Embley, Associate Editor

Accepted June 26, 2008