Evidence for Nucleomorph to Host Nucleus Gene Transfer: Light-Harvesting Complex Proteins from Cryptomonads and Chlorarachniophytes

James A. Deane, Martin Fraunholz, Vanessa Su, Uwe-G. Maier, William Martin, Dion G. Durnford, and Geoffrey I. McFadden

Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Parkville VIC 3052, Australia
Cell Biology and Applied Botany, Philipps-University Marburg, Karl-von-Frisch-Strasse, D-35032 Marburg, Germany
Institute of Botany III, University of Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany
Department of Biology, University of New Brunswick, Mail Service #45111, Fredericton, New Brunswick E3B 6E1, Canada

Submitted February 15, 2000; Accepted July 30, 2000
Monitoring Editor: Michael Melkonian

Cryptomonads and chlorarachniophytes acquired photosynthesis independently by engulfing and retaining eukaryotic algal cells. The nucleus of the engulfed cells (known as a nucleomorph) is much reduced and encodes only a handful of the numerous essential plastid proteins normally encoded by the nucleus of chloroplast-containing organisms. In cryptomonads and chlorarachniophytes these proteins are thought to be encoded by genes in the secondary host nucleus. Genes for these proteins were potentially transferred from the nucleomorph (symbiont nucleus) to the secondary host nucleus; nucleus to nucleus intracellular gene transfers. We isolated complementary DNA clones (cDNAs) for chlorophyll-binding proteins from a cryptomonad and a chlorarachniophyte. In each organism these genes reside in the secondary host nuclei, but phylogenetic evidence, and analysis of the targeting mechanisms, suggest the genes were initially in the respective nucleomorphs (symbiont nuclei). Implications for origins of secondary endosymbiotic algae are discussed.

Introduction

Cryptomonads and chlorarachniophytes are two groups of unicellular algae that have gained the ability to photosynthesise by capturing and maintaining a eukaryotic endosymbiont with a plastid (McFadden and Gilson 1995); a red alga in the case of cryptomonads (Douglas et al. 1991; Douglas and Penny 1999; Keeling et al. 1999; Maier et al. 1991; Van der Auwera et al. 1998) and a green alga in the case of chlorarachniophytes (Ishida et al. 1997; McFadden et al. 1995; Van de Peer et al. 1996). Although plastids are the most conspicuous feature of the cryptomonad and chlorarachniophyte endosymbionts, they also contain a miniature nucleus referred to as the nucleomorph (reviewed in McFadden and Gilson 1995.) The nucleomorph is the reduced nucleus of a red algal endosymbiont in cryptomonads and of a green algal endosymbiont in chlorarachniophytes.
Initially it was thought that nucleomorphs would encode numerous plastid proteins, similar to the nuclei of red and green algae (McFadden 1993). Indeed, this was the main rationale for the retention of nucleomorphs (McFadden 1993). However, sequence analyses reveal that cryptomonad and chlorarachniophyte nucleomorphs do not harbour numerous genes for plastid proteins. On the contrary, there is a distinct paucity of plastid protein genes and a surfeit of housekeeping genes in both nucleomorphs (Gilson et al. 1997; Gilson and McFadden 1996a; Gilson and McFadden 1996b; Zauner et al. 2000). The inescapable conclusion is that nucleomorph-encoded plastid protein genes have been transferred to the host cell nucleus. Here we investigate the fate of genes encoding cryptomonad and chlorarachniophyte plastid light-harvesting complex (LHC) proteins.

LHC proteins are essential for a functional photosynthetic plastid and are nuclear-encoded in both red algae and green algae, and indeed all algae thus far examined (Durnford et al. 1999; Green and Pichersky 1994). LHC genes should therefore have been present in the nuclei of the photosynthetic organisms that independently became the cryptomonad and chlorarachniophyte endosymbionts. We have isolated genes coding for cryptomonad and chlorarachniophyte LHC proteins and localised these genes to the host cell nuclei. Analysis of the N-terminal extensions on the predicted LHC proteins provides insight into how the proteins are targeted across the four membranes bounding the plastids of these algae. Additionally, phylogenetic analyses suggest that the LHC genes have undergone intracellular gene transfer from the nucleomorph to the host nucleus.

**Results**

**Cryptomonad LHC Clones**

Three cDNA clones (Lhcc4 = GenBank AF268322, Lhcc10 = GenBank AF268323, Lhcc13 = GenBank AF268324) were isolated from a Guillardia theta phage library. Lhcc13 is apparently a full length cDNA whereas Lhcc10 and Lhcc4 appear to be truncated at their 5' end and lack a start codon. Clones Lhcc10 and Lhcc13 were named Cac1 and Cac2 respectively in a previous publication (Durnford et al. 1999); the new names follow the consen sus nomenclature (Jansson et al. 1999).

Lhcc13 was investigated in detail as it was the longest cryptomonad LHC clone obtained and appeared to contain a complete coding sequence. Lhcc13 has an open reading frame of 717 bp and an apparently short 5' untranslated region. PCR-amplified genomic copies of Lhcc13 lacked introns and were identical in nucleotide sequence to the corresponding region of the Lhcc13 cDNA clone. Probing of a Northern blot with Lhcc13 labelled a transcript approximately 750 bases long (data not shown), which corresponds well with Lhcc13 clone size and suggests the cDNA is full length. Probing pulsed-field-separated Guillardia theta chromosomes with Lhcc13 gave a signal from nuclear chromosomes and no hybridisation to nucleomorph chromosomes or plastid DNA (Fig. 1).

The Lhcc13 predicted protein has an N-terminal hydrophobic region of 16 amino acids that is pre-
Lhcg11 (GenBank AF268320) and Lhcg12 (GenBank AF268321). Lhcg4 and Lhcg11 were previously named Lhc1 and Lhc2 respectively (Durnford et al. 1999). All three Chlorarachnion genes are very similar in nucleotide sequence and have open reading frames of approximately 1000bp. The predicted proteins differ mainly in the N-terminal regions (see below). PCR-amplified genomic copies of Lhcg12 lacked introns and were identical in nucleotide sequence to the corresponding region of the Lhcg12 cDNA clone. Northern blotting using Lhcg12 as a probe demonstrates the presence of 1000 base transcripts (data not shown).

Probing pulsed-field gels of Chlorarachnion sp. 621 chromosomal DNA with Lhcg12 gave a signal from nuclear chromosomes and no hybridisation to nucleomorph chromosomes (Fig. 4).

The Lhcg11 predicted protein begins with a hydrophobic region of 24 amino acids predicted to be a cleavable signal sequence by PSORT and SignalP to be a signal sequence cleaved between the alanine and phenylalanine at ASA|FMP (Fig. 2). A similar motif exists in Lhcc4 and Lhcc10, but because the clones are incomplete it is not possible to identify signal sequences at their N-termini. Downstream of the signal peptide of Lhcc13 is a region of about 24 amino acids (Fig. 2) with no obvious similarity to other LHC proteins but with some characteristics of plastid targeting transit peptides such as a lack of acidic amino acids and an elevated serine and threonine content (von Heijne 1991; von Heijne et al. 1991).

The cryptomonad LHC proteins each have three internal hydrophobic regions (Fig. 2) that correspond, in position and amino acid sequence, to the thylakoid spanning domains of LHC proteins from other eukaryotes (Durnford et al. 1999). All three cryptomonad Lhcc sequences branch within a clade that includes LHCs from red algae and algae containing chlorophylls a and c (Fig. 3). The three cryptomonad LHC proteins group with the red algal LHC proteins of Porphyridium cruentum and Cyanidium caldarium (Fig. 3).

Figure 2. Hydrophobicity plot of Lhcc13 predicted protein from Guillardia theta. A predicted signal peptide (S) is followed by a putative transit peptide (tp). Thereafter follows a region (MATURE PROTEIN) with three predicted membrane spanning helices with similarity to those of other LHC proteins. Basic residues are indicated with a ‘+’ sign, acidic residues with a ‘−’ sign, and hydroxylated residues with an ‘o’ sign.

Chlorarachnion sp. 621 LHC Clones
Three Chlorarachnion sp. 621 LHC cDNAs were recovered and designated Lhcg4 (GenBank AF268319), Lhcg11 (GenBank AF268320) and Lhcg12 (GenBank AF268321). Lhcg4 and Lhcg11 were previously named Lhc1 and Lhc2 respectively (Durnford et al. 1999). All three Chlorarachnion genes are very similar in nucleotide sequence and have open reading frames of approximately 1000bp. The predicted proteins differ mainly in the N-terminal regions (see below). PCR-amplified genomic copies of Lhcg12 lacked introns and were identical in nucleotide sequence to the corresponding region of the Lhcg12 cDNA clone. Northern blotting using Lhcg12 as a probe demonstrates the presence of 1000 base transcripts (data not shown). Probing pulsed-field gels of Chlorarachnion sp. 621 chromosomal DNA with Lhcg12 gave a signal from nuclear chromosomes and no hybridisation to nucleomorph chromosomes (Fig. 4).

The Lhcg11 predicted protein begins with a hydrophobic region of 24 amino acids predicted to be a cleavable signal sequence by PSORT and SignalP (Fig. 5). The Lhcg4 and Lhcg12 predicted proteins also have hydrophobic regions with possible cleavage sites near their N-termini, but in both predicted proteins this domain is preceded by an initial hydrophilic region of about 20 amino acid residues (not shown) that prevents PSORT and SignalP recognising them as signal sequences.

Downstream of the hydrophobic signal peptide of Chlorarachnion sp 621 Lhcg predicted proteins are
Figure 4. Location of LHC genes from Chlorarachnion sp. 621 by Southern blot analysis of pulsed-field gel separation of chromosomes. Lane 1 is Lambda ladder DNA standards. Lane 2 is chromosomal DNA from Chlorarachnion sp. 621. Nu = nuclear chromosomes; I, II and III = nucleomorph chromosomes one, two and three respectively. M = mitochondrial DNA. Lane 3 is a blot of lane 2 probed with Lhcg12.

Figure 5. Hydrophobicity plot of Chlorarachnion sp. 621 Lhcg11 predicted protein. A predicted signal peptide (SP) is followed by a putative transit peptide (tp). Thereafter follows a region (MATURE PROTEIN) with three predicted membrane spanning helices similar to those of other LHC proteins. Basic residues are indicated with a '+' sign, acidic residues with a '-' sign, and hydroxylated residues with an 'o' sign.
regions of approximately 50 amino acids with very few acidic residues and numerous hydroxylated residues (Fig 5). These regions exhibit no similarity to known chlorophyll a/b-binding proteins and are likely transit peptides. All three Lhcc proteins have more than double the proportion of serine residues in this region in comparison with downstream sequence (Lhcc4 has 11.5% serine in the transit peptide compared to 4.1% in the rest of the protein; Lhcg11 has 14.6% serine in the transit peptide compared to 4.5% in the rest of the protein; Lhcg12 has 8.5% serine in the transit peptide compared to 4.2% in the rest of the protein) which is reminiscent of plant and green algal transit peptides (von Heijne 1991; von Heijne et al. 1991). Analysis with ChloroP, a neural net trained to recognise transit peptides of plants, predicts that these regions are transit peptides that would be cleaved at the motif SPK|AN, which is shared by all the chlorarachniophyte Lhcg proteins identified thus far.

Chlorarachnion sp. 621 predicted Lhcg proteins have three internal hydrophobic domains (Fig. 5) typical of thylakoid spanning domains from LHC proteins of algae and plants that contain chlorophylls a and b. The three chlorarachniophyte LHC proteins differed by only a single amino acid in the predicted mature portion of the protein so only Lhcc4 was used in phylogenetic analyses. The phylogenetic trees indicate that chlorarachniophyte Lhcg proteins identified here are photosystem II-associated antenna proteins and are most closely related to Lhccb protein of the prasinophyte Tetraselmis (Fig 3).

Discussion

Cryptomonad and Chlorarachniophyte LHC Proteins

The cryptomonad LHC proteins contain three probable membrane-spanning hydrophobic regions with strong sequence similarity to conserved domains in all other eukaryotic chlorophyll-binding proteins (Durnford et al. 1999). The full length Lhcc13 clone from Guillardia theta encodes a precursor of 239 amino acids, which is comparable to LHC precursors from other chlorophyll a/c-containing algae and red algae (Apt et al. 1994; Apt et al. 1995a; Apt et al. 1995b; Bhaya and Grossman 1991; Caron et al. 1996; Durnford et al. 1996; Durnford and Green 1994; Hiller et al. 1995; Laroche et al. 1994; Tan et al. 1997). Phylogeny of the cryptomonad Lhcc proteins is consistent with a red algal origin for cryptomonad endosymbionts. In a previous analysis, (Durnford et al. 1999) reported that the proteins encoded by Lhcc10 and Lhcc13 undoubtedly belong to the clade including red algae and lineages containing chlorophylls a and c. In that study, only a weak relationship between cryptomonad endosymbiotic Lhcc10 and Lhcc13 proteins and LHC proteins currently known from red algae was observed (Durnford et al. 1999). Here we demonstrate a similar (albeit still weak) relationship in trees that contain a new cryptomonad LHC (Lhcc4) and a new red algal LHC (Fig. 3). This further confirms the red algal ancestry of the cryptomonad endosymbiont. Divergence of cryptomonad and red algal LHC proteins probably reflects the different antenna configurations. Red algae utilise chlorophyll a and phycobilisomes but cryptomonads utilise chlorophylls a and c2. Furthermore, the phycobilins of cryptomonads are not organised into phycobilisomes but are soluble and located in the thylakoid lumen (Gantt et al. 1971; Ingram and Hiller 1983; Ludwig and Gibbs 1989; Spear-Bernstein and Miller 1985).

Chlorarachniophyte LHC proteins also contain three probable membrane spanning domains and precursors are in the order of 350 amino acids long, which is similar in size to those of green algal and plant LHC precursors (Green and Kühlbrandt 1995). As previously reported (Durnford et al. 1999), LHC phylogeny confirms a green algal source for the chlorarachniophyte endosymbiont and predicts that the chlorarachniophyte LHCs reported here are part of photosystem II. The analysis reported here shows a strong relationship between chlorarachniophyte LHC protein and that of the prasinophyte Tetraselmis sp. (Fig. 3), which is consistent with pigment data (Sasa et al. 1992). Other phylogenetic analyses have indicated that the chlorarachniophyte endosymbiont is related to ulvophytes (Ishida et al. 1997) so it will be interesting to compare ulvophyte LHC proteins when they become available.

Targeting of LHC Proteins to the Cryptophyte and Chlorarachniophyte Plastids

The nucleus-encoded LHC genes of cryptomonads and chlorarachniophytes are almost certainly translated on ribosomes in the host cell cytoplasm. Consequently, they must cross four membranes before insertion into the thylakoid membranes. This passage is probably mediated by N-terminal extensions on the encoded proteins. Cryptomonad LHC proteins have a bipartite presequence that appears to comprise of a signal sequence followed by a plastid transit peptide (Fig 2). Lhcc13 and Lhcc4 amino acid sequences both have a GPSM|IQ motif similar to a putative cleavage site for red algal plastid transit
peptides (Fig 6). A similar motif (GPXM|Q) has also been noted in the presequences of nuclear encoded Guillardia theta and Pyrenomonas salina GapC1 (chloroplast glyceraldehyde phosphate dehydrogenase) proteins (Liaud et al. 1997), and a more divergent motif (KMTM|Q) occurs in Lhcc10 (Fig 6). No similar motif is apparent in the transit peptides of the two nucleomorph-encoded, plastid-targeted proteins (FtsZ and rubredoxin) of cryptomonad proteins characterised thus far (Zauner et al. 2000), but this may reflect the extreme divergence of nucleomorph encoded proteins (Keeling et al. 1999).

Chlorarachniophyte LHC proteins have presequences that also contain possible signal sequences followed by transit peptide regions (Fig 5). However, Lhcg4 and Lhcg12 proteins also have an initial N-terminal hydrophilic region (not shown) not typically found at the start of signal sequences. There is no obvious function for this hydrophilic region and it is difficult to say how it might affect a signal sequence it preceded. The chlorarachniophyte LHC proteins also contain a putative transit peptide according to neural net analysis (Fig 5).

A bipartite (signal peptide/transit peptide) presequence configuration, as apparent in cryptomonad and chlorarachniophyte, is hypothesised to allow signal sequence-mediated transport and entry to the endomembrane compartment housing the endosymbiont, followed by transit peptide-mediated transport into the plastid (McFadden 1999). This model has been proposed for plastid targeting of cryptomonad nuclear-encoded plastid GAPDH (Liaud et al. 1997) and phycoerythrin (McFadden and Gilson 1997; Reith 1995), as well as a range of nuclear-encoded plastid proteins from other organisms with presequences followed by transit peptides surrounded by more membranes (for reviews see (Bodyl 1997; Bodyl 1999a; Bodyl 1999b; Lang et al. 1998; McFadden 1999; Schwartzbach et al. 1998]).

Cryptophyte nuclear-encoded plastid GAPDH (Liaud et al. 1997) have bipartite presequences with sequence identity to those of cryptophyte LHC precursors (Fig 6). Bipartite presequences have also been described for diatom fucoxanthin-chlorophyll proteins and nuclear-encoded plastid ATPase gamma subunits (Apt et al. 1993; Bhaya and Grossman 1991; Kroth-Pancic 1995; Lang et al. 1998; Pancic and Strotmann 1993), brown algal fucoxanthin-chlorophyll proteins (Apt et al. 1995a), a variety nuclear encoded euglenophyte plastid proteins (Chan et al. 1990; Henze et al. 1995; Lin et al. 1994; Sharif et al. 1989; Shashidhara et al. 1992; Sulli et al. 1999; Sulli and Schwartzbach 1995; Sulli and Schwartzbach 1996), and nuclear-encoded apicomplexan plastid proteins (Waller et al. 2000; Waller 1998). Several lines of evidence demonstrate that the signal peptide components of diatom (Bhya and Grossman 1991), cryptomonad (Wastl and Maier 2000) and apicomplexan plastid leaders (Waller et al. 2000) target proteins to plastids in the first instance via cotranslational insertion into the endomembrane system. It has also been shown that the second, transit peptide-like component of the leader from diatoms (Lang et al. 1998) and cryptomonads (Wastl and Maier 2000) is able to effect import into isolated plant chloroplasts in vitro. Dinoflagellate nuclear-encoded, peridinin-chlorophyll proteins have presequences that often appear tripartite (Norris and Miller 1994; Sharples et al. 1996). Nevertheless, these proteins begin with a hydrophobic domain that is probably a signal sequence targeting them to the endoplasmic reticulum (Norris and Miller 1994; Sharples et al. 1996). Thus, an initial endoplasmic reticulum targeting step seems to be a common feature of protein transport into plastids enclosed by three (dinoflagellates, euglenophytes) or four membranes (diatoms, brown algae, apicomplexa, cryptomonads and chlorarachniophytes).

It is interesting to compare how signal sequences might direct LHC proteins into the cryptophyte and chlorarachniophyte endosymbionts. The outer membrane surrounding the cryptomonad endosymbiont is continuous with the endoplasmic reticulum and covered with ribosomes, and, as first suggested by Gibbs, cotranslational insertion of LHC proteins produced on the outer endosymbiont membrane would provide the most direct route into the cryptomonad endosymbiont (Gibbs 1979; Gibbs 1981; Hofmann et al. 1994). However, unlike cryptomonads, the outer membrane of the chlorarachniophyte endosymbiont is not continuous with the endoplasmic reticulum and does not bear ribosomes (Hibberd and Norris 1984). This means that the cotranslational model of protein import proposed for cryptomonads cannot operate for chlorarachniophytes (Bodyl 1997; McFadden and Gilson 1997). Chlorarachniophyte LHC proteins may enter the host cell endomembrane system and travel to the endosymbiont in vesicles, either via the Golgi (Bodyl 1997), as demonstrated for Euglena plastid proteins (Osafune et al. 1991; Sulli et al. 1999), or directly from the endoplasmic reticulum.

In both models a signal sequence mediates initial entry into the endosymbiont containing compartment, but how the proteins pass from the endomembrane lumen into the endosymbionts remains mysterious (McFadden 1999). Signal sequences usually insert proteins through a single membrane rather than the paired outer membranes that bound the endosymbionts. Hence, the bipartite model can ac-
count for the outer membrane (signal sequence) and the inner pair of membranes (transit peptide), but provides no explanation for transport across the membrane thought to represent the endosymbiotic plasma membrane (McFadden 1999). One suggestion is that this membrane has a set of channels homologous to those of the plant outer chloroplast membrane (the so-called Toc apparatus) that acts in tandem with similar apparatus in the subtending membrane (third from the outside) (Cavalier-Smith 1999; van Dooren et al. 2000). A competing hypothesis suggests that vesicles form from the former endosymbiont membrane to encapsulate plastid-bound proteins and deliver them across the subtending membrane (third from the outside) by fusion with it (Kroth and Strotmann 1999). Yet another option is that the membrane is highly porous (Kroth and Strotmann 1999), but this model would not work in cryptomonads and chlorarachniophytes because the endosymbiont cytoplasmic proteins would ‘leak out’ into the endomembrane system lumen and be in danger of being secreted. There is currently no compelling evidence to support any model, and none of the models rationalises the apparent redundancy of this particular membrane (van Dooren et al. 2000).

Intracellular Gene Transfer

Phylogenetic analyses demonstrate that the nuclear LHC genes in cryptomonads and chlorarachniophytes most probably derive from their respective endosymbionts (red and green algae). Two possible sources of the genes are the plastid genome or the endosymbiont nucleus (nucleomorph). Various lines of evidence support a nucleomorph source for these genes.

Thus far, all known genes for three helix LHC proteins are nuclear encoded (Durnford et al. 1999), and it is postulated that they evolved from single helix plastid-encoded proteins by duplication after transfer to the nucleus (Durnford et al. 1999; Green and Pichersky 1994). Complete plastid genome sequences are available for red algae and green algae and none encode a three helix LHC. This strongly suggests that the LHC genes in the cryptomonad and chlorarachniophyte endosymbionts would have been housed in the respective endosymbiont nuclei at the outset of the two secondary endosymbioses. This is in contrast to proposals that the hosts were already in possession of suitable genes for plastid proteins previously acquired from hypothetical prokaryotic endosymbionts that the new, secondary symbionts replaced (Bodily 1999a; Häuber et al. 1994).

Further evidence for nucleomorph to nucleus gene transfer comes from the comparison of transit peptides of red algae and cryptomonad LHCs. If our assertion that the cryptomonad LHC genes ultimately derive from the nucleomorph is correct, the protein would have originally been targeted to the plastid in the free living alga by a transit peptide similar to those of modern red algal plastid-targeted proteins. After secondary endosymbiosis, relocation of such a gene to the secondary host nucleus could easily have included this transit peptide, especially since it would probably be necessary in order for the protein to be targeted back to the plastid (McFadden 1999). The similarity, particularly in the putative cleavage motif (GPXM|XX), between cryptomonad and red algal transit peptides (Fig 6) strongly suggests that this is exactly what happened; the transit peptide region and mature protein region were relocated from the nucleomorph to the nucleus simultaneously. Thus, just as conservative sorting of proteins within primary plastids demonstrates the reuse of thylakoid targeting motifs in plastid to nucleus gene transfers (Yuan et al. 1994), our characterisation of cryptomonad LHC leaders suggests that a similar recycling of targeting motifs (in this case transit peptides) occurs in secondary endosymbiotic gene relocation. In this respect each targeting step (leader motifs) can be regarded as a legacy of intracellular gene transfers, the encoded protein tracing a path back through its previous homes.

Interestingly, neither of the LHC genes examined at the genomic level (cryptomonad Lhcc13 and chlorarachniophyte Lhcg12) contained introns. Absence of introns in the cryptomonad Lhcc13 gene is consistent with the relative scarcity of spliceosomal introns in cryptomonad nucleomorph genes (Zauner et al. 2000). Lack of introns in chlorarachniophyte Lhcg12 is more intriguing.

Chlorarachniophyte nucleomorph genes typically have numerous introns (Gilson et al. 1997; Gilson and McFadden 1996a; Gilson and McFadden 1996b) so it might be expected that a transferred gene would carry some legacy (introns) of this ancestry. We are not able to establish whether or not Lhcg 12 contained introns prior to transfer from the nucleomorph to the host nucleus in Chlorarachnion sp 621, but the lack of introns perhaps suggests that it never contained any. This may be germane to the proposal that introns (which are highly unusual in the chlorarachniophyte nucleomorph) might be an obstacle to nucleomorph to host nucleus transfer (Gilson et al. 1997). Perhaps only intronless genes were successfully transferred. Alternatively, transfer might have involved an RNA intermediate from which the introns had been excised. It will be interesting to examine the intron/exon architecture of other transferred genes.
**Figure 6.** Alignment of the leaders from selected nuclear encoded plastid proteins of heterokonts (upper block), cryptomonads (middle block) and red algae (lower block). Signal peptides are in upper case, transit peptides lower case and mature proteins in upper case. The cleavage site of the transit peptide has only been experimentally characterised in *Odontella sinensis* prk and *Phaeodactylum tricornutum* atpC. **Odontella sinensis** prk, Laminaria saccharina Lhcf1, Macroystis pyrifera Lhcf1, Heterosigma carterae Fcp1, Skeletonema costatum scfcp8, Cyclotella cryptica cpC2, Isochrysis galbana Fcp1, Giraudyopsis stellaris fcp, Cyclotella ACCase, Guillardia theta Lhcc10, Guillardia theta Lhcc13, Guillardia theta Lhcc4, Guillardia theta GapC1, Prymnesiopsis salina GapC1, Chondrus crispus GapA, Gracilaria verrucosa GapA, Porphyridium cruentum lhcA, Aglaosthionon neglectum R-pe gamma subunit, Aglaosthionon neglectum 31 kDa pe subunit, Cyanidium caldarium lhcA.
Evolutionary Considerations

Intriguingly, the signal peptide cleavage sites for known cryptophyte and heterokont plastid proteins bear close resemblance, often being AS/AA|FXP (Fig. 6). This similarity is potentially of importance to proposals that cryptophytes and heterokonts derive from the same secondary endosymbiosis, the so-called Kingdom Chromista hypothesis (Cavalier-Smith 1989; Cavalier-Smith et al. 1994; Cavalier-Smith et al. 1996). The similarity of cryptophyte and heterokont signal sequences could be explained if the bipartite plastid targeting machinery of both groups first evolved in an ancestral lineage that diverged to give rise to cryptophytes and heterokonts, as argued by Cavalier-Smith (Cavalier-Smith 1989; Cavalier-Smith et al. 1994; Cavalier-Smith et al. 1996). However, the observed similarity must be viewed with caution. The difficulty here is distinguishing between common origin and convergence. Functional constraints dictate that signal sequences must be similar in all eukaryotes; the requirement for hydrophobicity and a signal peptidase cleavage consensus (von Heijne 1985; von Heijne 1999) make amino acid content much alike in all signal sequences (Nielsen et al. 1997). Thus, signal sequence similarity could be convergence rather than an indication of common ancestry. Nevertheless, the similarity we observe extends at least three amino acids downstream of the signal peptide cleavage site (Fig 6), and this region is not known to be important for signal sequence function (Nielsen et al. 1997; von Heijne 1999). This tempts us to speculate that FXP (or similar domains) at the beginning of the transit peptide may have some significance to plastid targeting. In this respect it is interesting that numerous proteins targeted to the apicomplexan plastid, which is surrounded by four membranes, also have a phenylalanine (F) at the +1 position after the signal peptide cleavage site (Waller et al. 1998). Phenylalanine (F) is infrequent at this site in other (secreted) eukaryotic proteins (only 5.1% of secreted proteins have a phenylalanine at +1 (Nielsen et al. 1997)), further implicating this residue as important in secondary endosymbiont targeting.

Returning to the original topic of whether or not cryptophytes and heterokonts derive from a common secondary endosymbiosis, we note that cryptomonad plastid proteins clearly have red algal-like transit peptide cleavage motifs (Fig. 6), which is in accord with a red algal origin for the endosymbiont. Comparison of the putative transit peptides of cryptomonads and red algae with those of heterokont algae also reveals possible similarities (Fig. 6).

Lang et al (Lang et al. 1998) have suggested that a methionine residue often defines the C-terminal end of ‘chromophyte’ transit peptides, and we extend this similarity to include red algal and cryptomonad transit peptides (Fig 6). If the plastid targeting machinery of cryptophytes and heterokonts had a common origin (Cavalier-Smith 1989), we might also expect to see homology between their transit peptides, particularly the cleavage sites, and it will be interesting to gain more experimental evidence to verify or refute the predictions presented in Fig 6.

At present we do not feel that the data either refute or confirm the Kingdom Chromista hypothesis. The issue is complicated by the fact that heterokont plastids are also believed to originate from red algal endosymbionts (Bhattacharya and Medlin 1995) so red algal like transit peptides are to be expected on plastid targeted proteins of heterokonts. Similarity between transit peptides of cryptomonads and heterokonts with red algae need not necessarily indicate a common secondary endosymbiosis as postulated in the Kingdom Chromista hypothesis. Just as it has been necessary to define shared derived characteristics for all chloroplasts to infer single primary endosymbiosis (Stoebe and Kowallik 1999), it will also be necessary to define shared derived characters for plastid targeting in secondary endosymbionts to infer homology.

Concluding Remarks

The LHC genes of cryptomonads and chlorarachniophytes have very probably undergone nucleomorph to nucleus intracellular gene transfer. The driver for this transfer was most likely the loss of recombinational opportunity and the inherent genetic bottleneck in the nucleomorph when it became an endosymbiont nucleus (McFadden 1999). The cognate loss of ability to recover from deleterious mutations likely favoured transfer of genes to the host nucleus (McFadden 1999).

Ultimately, LHC genes are believed to have originated from single helix encoding plastid genes that relocated to the nucleus early in primary endosymbiosis (Durnford et al. 1999; Green and Pichersky 1994). The chlorarachniophyte and cryptomonad LHC genes thus seem to be genetic vagabonds that have undergone two sets of intracellular gene transfer, one from the plastid to the nucleus in primary endosymbiosis, then a second transfer from this nucleus (now a nucleomorph) to the host nucleus in secondary endosymbiosis. The bipartite targeting leader system would seem to be a reverse recapitulation of these transfers leading the product back to its source compartment.
Methods

Clones of LHC genes from a cryptomonad (Guillar-dia theta; CCMP strain #325) and a chlorarachniophyte (Chlorarachnion sp. 621; CCMP strain #621) were obtained from Lambda ZAP (Stratagene) cDNA libraries (Gilson and McFadden 1996b) by screening with a 305bp Apai/Bglll fragment of Lemna gibba Lhc gene (GenBank M12152). An \([\alpha^{32P}]\) dCTP-labelled Lemna gibba Lhc probe was prepared using Megaprime (Amersham). Phage clones containing LHC cDNAs were auto-excised as a pBluescript SK phagemid using the manufacturer's protocol. DNA sequencing was done with Dye Terminator chemistry (Perkin-Elmer) and an ABI 373a automated sequencer (Perkin-Elmer). Contig assembly was done with Sequencher 3.0 (Gene Codes).

Additional Chlorarachnion sp. 621 cDNA clones were obtained by PCR amplification from phage DNA prepared from the Chlorarachnion sp. 621 cDNA library as previously described (Keeling et al. 1998). M13 reverse primer was used with an internal primer \((5'\text{CACGCAGACCTTGACGAGC3'})\) designed to match LHC cDNAs obtained by library screening. PCR conditions were 15 cycles of 1min at 95°C, 1min at 35°C, 1min at 72°C then 15 cycles of 1min at 95°C, 1min at 50°C, 1min at 72°C. PCR products were purified using Prep-A-Gene (BioRad) and cloned using pGEM T-vector (Promega).

To amplify genomic clones, primers were designed to match regions at the 5’ and 3’ end of G. theta Lhc13 and Chlorarachnion sp. 621 Lhcg12. Lhcc13 primers were \(5'TG GCCATGTTCCGAGCTG3'\) & \(5'AT TCGCCATGTTCCGAGCTG3'\) and \(5'\text{CAGCGCCGACCTTGACGAGC3'}\) designed to match regions at the 5’ and 3’ end of Lhc gene (GenBank M12152). An \([\alpha^{32P}]\) dCTP-labelled Lemna gibba Lhc probe was prepared using Megaprime (Amersham). Phage clones containing LHC cDNAs were auto-excised as a pBluescript SK phagemid using the manufacturer's protocol. DNA sequencing was done with Dye Terminator chemistry (Perkin-Elmer) and an ABI 373a automated sequencer (Perkin-Elmer). Contig assembly was done with Sequencher 3.0 (Gene Codes).

Additional Chlorarachnion sp. 621 cDNA clones were obtained by PCR amplification from phage DNA prepared from the Chlorarachnion sp. 621 cDNA library as previously described (Keeling et al. 1998). M13 reverse primer was used with an internal primer \((5'\text{CACGCAGACCTTGACGAGC3'})\) designed to match LHC cDNAs obtained by library screening. PCR conditions were 15 cycles of 1min at 95°C, 1min at 35°C, 1min at 72°C then 15 cycles of 1min at 95°C, 1min at 50°C, 1min at 72°C. PCR products were purified using Prep-A-Gene (BioRad) and cloned using pGEM T-vector (Promega).

To amplify genomic clones, primers were designed to match regions at the 5’ and 3’ end of G. theta Lhc13 and Chlorarachnion sp. 621 Lhcg12. Lhcc13 primers were \(5'TG GCCATGTTCCGAGCTG3'\) & \(5'\text{CAGCGCCGACCTTGACGAGC3'}\). Lhcg12 primers were \(5'\text{GAGCTCCCTCCTCAACGCTAC3'}\) and \(5'\text{TGTGAAGTTGGGGATCCGG3'}\). Lhcg12 primers were \(5'\text{CAGCGCCGACCTTGACGAGC3'}\) designed to match LHC cDNAs obtained by library screening. PCR conditions were 15 cycles of 1min at 95°C, 1min at 35°C, 1min at 72°C then 15 cycles of 1min at 95°C, 1min at 50°C, 1min at 72°C. PCR products were purified using Prep-A-Gene (BioRad) and cloned using pGEM T-vector (Promega).


Northern blots of isolated RNA and Southern blotting of pulsed-field gels were performed as previously described (Gilson and McFadden 1996b). Predicted protein sequences of Guillar-dia theta and Chlorarachnion sp. 621 LHC clones were aligned to other LHC proteins as previously described (Durnford et al. 1999). Phylogenetic trees were constructed using PAUP* 4.02b (Swofford 2000). Bootstrap values for neighbor-joining distance and maximum parsimony are for 100 replicates. Quartet puzzling trees were constructed using 1000 puzzling steps. The tree was rooted with the Euglena gracilis Lhca35 & 38 (Durnford et al. 1999). The alignment is available on request. Accession numbers are given in (Durnford et al. 1999) except for Cyanidium caldarium (A 012759) and Tetraselmis sp. RG-15 (AF017998).

References


Gibbs SP (1979) The route of entry of cytoplasmically synthesized proteins into chloroplasts of algae possessing chloroplast ER. J Cell Sci 35: 253–266


Gilson PR, McFadden GI (1996b) The miniaturised nuclear genome of a eukaryotic endosymbiont contains genes that overlap, genes that are transcribed, and smallest known spliceosomal introns. Proc Natl Acad Sci USA 93: 7737–7742


Tan S, Ducret A, Aebersold R, Gantt E (1997) Red algal LHC I genes have similarities with both chl a/b- and chl a/c-binding proteins – a 21 kDa polypeptide encoded by lhcaR2 is one of the six LHC I polypeptides. Photosynth Res 53: 129–140


