Photosynthesis Research **39**: 149–162, 1994.

Regular paper

Hypothesis for the evolution of three-helix Chl *a*/*b* and Chl *a*/*c* light-harvesting antenna proteins from two-helix and four-helix ancestors

Beverley R. Green[1] & Eran Pichersky[2]
[1]*Botany Department, University of British Columbia, Vancouver, B.C. Canada V6T 1Z4*; [2]*Biology Department University of Michigan, Ann Arbor, MI 48109, USA*

Received 10 February 1993; accepted in revised form 25 October 1993

Key words: protein family, *psb*S protein, algae, gene duplication, LHC I, LHC II, CP29

Abstract

The nuclear-encoded Chl *a*/*b* and Chl *a*/*c* antenna proteins of photosynthetic eukaryotes are part of an extended family of proteins that also includes the early light-induced proteins (ELIPs) and the 22 kDa intrinsic protein of PS II (encoded by *psb*S gene). All members of this family have three trans-membrane helices except for the *psb*S protein, which has four. The amino acid sequences of these proteins are compared and related to the three-dimensional structure of pea LHC II Type I (Kühlbrandt and Wang, Nature 350: 130–134, 1991). The similarity of *psb*S to the three-helix members of the family suggests that the latter arose from a four-helix ancestor that lost its C-terminal helix by deletion. Strong internal similarity between the two halves of the *psb*S protein suggests that it in turn arose as the result of the duplication of a gene encoding a two-helix protein. Since *psb*S is reported to be present in at least one cyanobacterium, the ancestral four-helix protein may have been present prior to the endosymbiotic event or events that gave rise to the photosynthetic eukaryotes. The Chl *a*/*b* and Chl *a*/*c* antenna proteins, and the immunologically-related proteins in the rhodophytes may have had a common ancestor which was present in the early photosynthetic eukaryotes, and predated their division into rhodophyte, chromophyte and chlorophyte lineages. The LHC I–LHC II divergence probably occurred before the separation of higher plants from chlorophyte algae and euglenophytes, and the different Types of LHC I and LHC II proteins arose prior to the separation of angiosperms and gymnosperms.

Abbreviations: CAB – Chl *a*/*b*-binding; ELIP – early light-induced protein; FCP – fucoxanthin-Chl *a*/*c* protein; PCR – polymerase chain reaction; TMH – trans-membrane helix

Introduction

The Chl *a*/*b*-binding proteins (CABs) of higher plant light-harvesting complexes are part of an extended family of proteins (Green et al. 1991). This family includes the fucoxanthin-Chl *a*/*c* antenna proteins (FCPs) of chromophyte algae, and the Early Light-Induced Proteins (ELIPs) of higher plants. Hydropathy plots of all family members predict three trans-membrane helices (TMHs). All the proteins share two highly conserved regions comprising the first and third TMHs and the stroma-exposed, high beta-turn regions preceding them. The two conserved regions share considerable sequence similarity, indicating that they arose as a result of a gene duplication (Hoffmann et al. 1987, Pichersky and Green 1990).

The three-dimensional structure of pea LHC II Type I has been determined at 6 Å resolution

(Kühlbrandt and Wang 1991), and analysis of 3 Å data is in progress (Kühlbrandt and Wang 1992). The 6 Å structure showed that the Chl-protein does have the three TMHs predicted from hydropathy plots. The two conserved helices cross each other at an angle of about 56° (31° and 25° from membrane normal) and project somewhat above the plane of the membrane, while the middle helix is shorter and almost perpendicular to the membrane plane. The conserved beta-turn regions appear as curving arms on the surface of the membrane. Thirteen or fourteen Chl molecules are bound in two layers around and between the helices, with their rings oriented approximately perpendicular to the membrane plane. Because of the high degree of sequence conservation in the family, it is reasonable to conclude that all its members will have the same overall fold (Green et al. 1992a).

We have recently isolated a cDNA clone for *psb*S, the gene encoding the 22 kDa intrinsic protein of PS II, which does not bind detectable amounts of Chl (Kim et al. 1992). Sequence comparisons show that it is also a member of the CAB family. However, this protein has four, rather than three TMHs. The third helix is related to the first helix, as in the CABs, and the fourth helix is related to the second helix. In other words, the protein appears to be the result of the duplication of a gene encoding a two-helix protein. This suggests that the entire family of three-helix proteins may have arisen from a two-helix protein that was duplicated to give a four-helix protein, and subsequent loss of most of the fourth TMH. In this paper, we will first compare the amino acid sequences of the extant three-helix members of this family (the CABs, FCPs and ELIPs) and then discuss the evolutionary implications of the *psb*S sequence.

Materials and methods

All members of this extended family share enough sequence similarity so that most of the alignment can be done manually, taking care to superimpose equivalent secondary structures where possible and to keep gaps to a minimum. Similar amino acids were defined using a modification of the amino acid groupings defined by Taylor (1986): specifically FILMV (aliphatic non-polar), FYW (aromatic), AG (tiny). Also considered as similar were DE (acidic), ND and QE (acid and amide), ST (hydroxylated) and HNQ (known or possible Chl ligands). Structure prediction (Green 1990) was done using several different hydropathy scales, two window sizes, the beta-turn scales of Wilmot and Thornton (1988) and the helix-end propensities of Richardson and Richardson (1988) The first and third membrane-spanning helices were extended on the N-terminal (stromal) side using alpha-helical propensities derived from soluble proteins (Chou and Fasman 1974) because the electron diffraction structure showed the helices projecting about 8 Å (2–3 turns) above the non-polar core of the bilayer (Kühlbrandt and Wang 1992).

We found that standard multiple alignment programs (Doolittle, Eugene package or Pileup in Wisconsin GCG package) could not cope with the very different lengths of the middle sections of the CAB proteins and could not recognize the characteristic pattern of conserved residues in the C-terminal half of the second TMH which is very evident on visual examination. This motif was recognized by the program MACAW (Schuler et al. 1991), in which blocks of sequence having significant similarity are aligned. MACAW was used to test the visual alignment and to improve it in several sections (e.g. N-terminal region) where there was little obvious similarity.

Methods used in gene cloning and sequencing, and the identification of genes with the proteins they encode are given in our published work (refs. in Green et al. 1992b).

Results

The CAB (Chl a/b) protein family of tomato

Higher plants have four Chl *a/b* protein complexes: LHC I which is associated with PS I, and LHC II, CP 29/CP 26 and CP 24 associated primarily with PS II. These complexes have four, three, two and one unique polypeptide types, respectively (Green et al. 1991). Genes of the

same type encode almost identical amino acid sequences and have the same number and position of introns (Chitnis and Thornber 1988). We have cloned and sequenced all but one of the members of this diverse family of Chl *a*/*b*-proteins in tomato, and linked many of the genes to the polypeptides they encode by tryptic peptide sequencing (Green et al. 1992b, Pichersky et al. 1989, 1991, Schwartz and Pichersky 1990, Schwartz et al. 1991 and references therein).

The availability of this array of sequences (Fig. 1) allows us to draw some general conclusions about the CAB proteins and to relate them to the three-dimensional structure of pea LHC II Type I (Kühlbrandt and Wang 1992). In Fig. 1 the deduced protein sequences of the four LHC I polypeptide Types are given first, followed by the three LHC II Types, CP 29 Type I (also called CP 26 or CP 27) and CP 24. In most of the figures in this paper, the sequences are broken up into four blocks for reasons of space: they will be referred to as the N-terminal region, the first conserved region, the middle region, and the second conserved region. The diagram below shows the relationship between the four regions and the secondary structure elements. The insert indicates the overall topology of the molecule. The first and third TMHs are predicted to include a number of polar residues at the N-terminal end because the 6 Å structure showed that these helices were 31–33 amino acids long and projected above the membrane plane (Kühlbrandt and Wang 1992).

Figure 1 shows the high degree of sequence similarity in the two conserved regions of all the tomato CABs. The relatedness of the two conserved regions to each other (Hoffmann et al. 1987) can be seen by comparing the starred residues in both regions. The high degree of sequence conservation starts N-terminal to the first and third TMH, where the sequence has a high beta-turn probability (Wilmot and Thornton 1988) and corresponds to two curving 'arms' on the surface of the protein (Kühlbrandt and Wang 1992). In these arms there are a number of highly conserved or conservatively substituted motifs, especially the FDPLGL motif (F replaced by W or Y in several cases). There are also conserved G, P, D and L residues. At other positions the character of the side-chain is conserved, e.g. A or G (tiny), I, L, M, V or F (aliphatic hydrophobic), F, Y or W (aromatic) (Taylor 1986).

In the first and third TMHs there are a number of strikingly conserved residues (stars) which are also conserved in other sections of the extended family. Both helices have a His (H) or Asn (N) followed by a small amino acid (Gly (G), Ala (A) or Cys (C)), then an absolutely conserved Arg (R). They are followed by W/FAML or LAML/F motifs. In the first TMH there are two highly conserved Glu (E), one of which is replaced by Gln (Q) in CP 24 apoprotein. The analogous positions in the third TMH are occupied by E and Q (E in CP 24). A helical wheel plot (Schiffer and Edmundson 1967) showed that the sidechains of the E–E or E–Q pairs project on the same side of the helix (data not shown). H, N and Q are known to be capable of forming ligands to the Mg atom of Chl from studies on the purple bacterial reaction centre (Coleman and Youvan 1990). Since E and Q seem to be interchangeable in these positions, it is conceivable that the glutamate carbonyl is also a Chl fifth ligand; alternatively it could be H-bonding to one of the polar substituents on the Chl ring. However, since the E–E and E–Q pairs are found in all members of the extended family, even those that do not bind Chl (see below), they could perform some vital function in assembly or membrane insertion.

The last highly conserved segment is just C-terminal to the third TMH, outside the hydrophobic region. It is predicted to be an amphipathic alpha-helix (Green et al. 1991). There is a moment of conservation as well as a hydrophobic moment, and the two moments are orthogonal (data not shown). This implies that the most highly conserved residues are at the interface between the polar and non-polar milieus. Among them are several potential Chl ligands: H, N and possibly E or D. It may be significant that this 'tail' is missing in the apoprotein of the notoriously unstable CP 24 complex, which also has a very low Chl *a*/*b* ratio (Dunahay and Staehelin 1986).

Much of the differentiation among CAB proteins appears to reside in the central helix and its flanking regions (Schwartz et al. 1991, Green et al. 1992a). This region was first aligned by

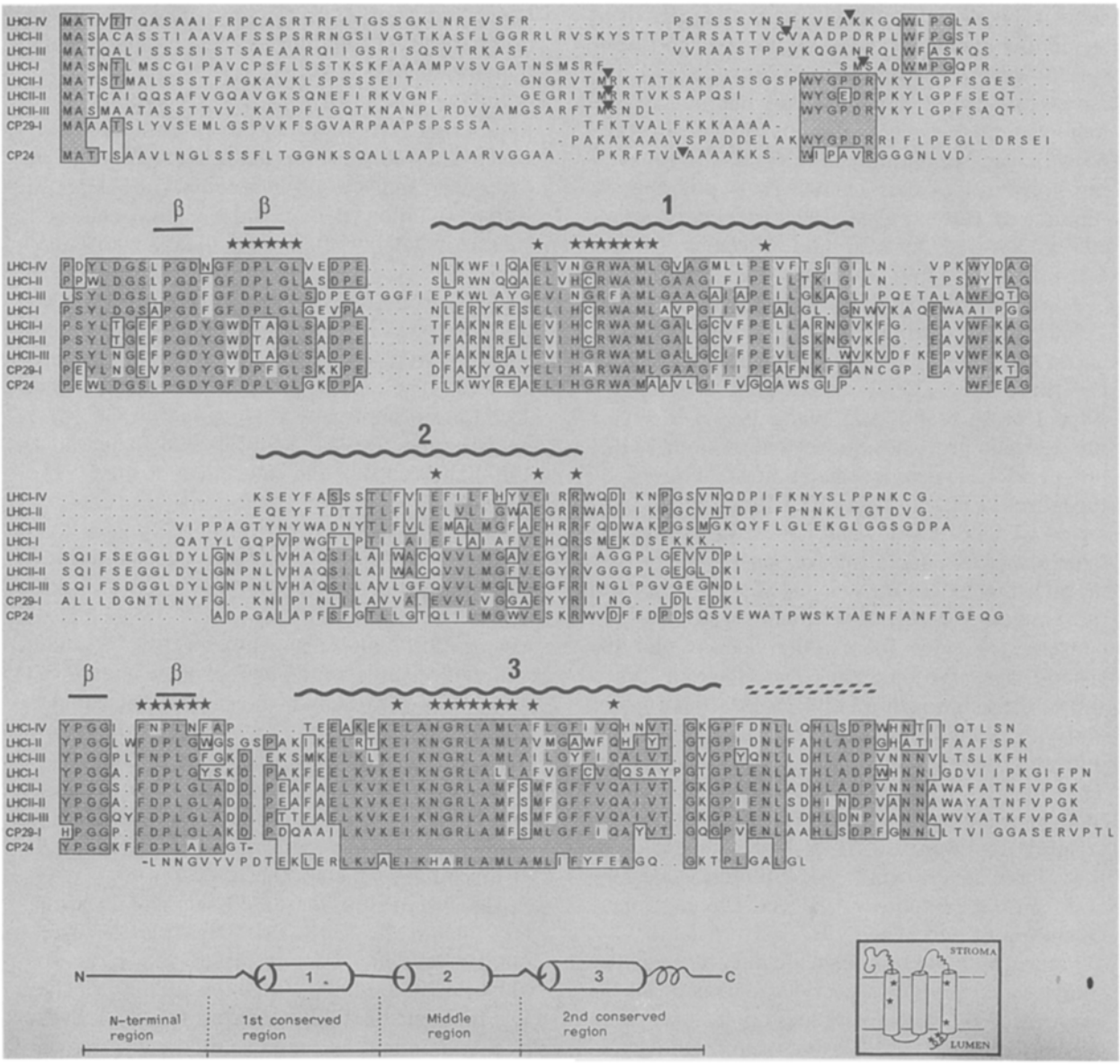

Fig. *1.* Amino acid sequences of the tomato CAB polypeptides. Shaded residues: identical in at least four proteins (five if includes all LHC Is or all LHC **IIs**, except for the PS I and PS II specific motifs in the N-terminal region) of tomato; boxed residues: similar (conservative substitutions ILMVF, FYW, AG, EQ, DN, ED, HN, RK); star: residue discussed in text; heavy boxes: PS I and PS II N-terminal region motifs; arrowhead: transit peptide cleavage site; period: gaps inserted to align sequences; –: sequence broken into two lines; heavy wavy arrow: TMH; beta: predicted beta turn; double hatched line: predicted amphipathic helix. Diagram (bottom) shows the division of the sequences into four regions. Insert: Topology of LHC II-I.

superimposing hydropathy plots and helix end predictions (Green 1990). This brought to our attention that in all the CABs there are two conserved E or E/Q residues, and a conserved R that probably acts as a stop-transfer signal at the end of the helix. The multiple alignment program MACAW (Schuler et al. 1991) also recognized this region as having significant similarity. Three of the LHC I proteins have considerably longer connectors between the second and third helices than do LHC II and CP29 Type I. Conversely, the latter have longer connectors

between the first and second helices. However, the differences in molecular weights of the mature proteins are mainly the result of the lengths of the mature N-terminal region. Although CP 24 would be expected to be more like LHC II, its connectors are more like those of LHC I. All the flanking connectors are highly polar, suggesting they form surface-exposed loops. This in turn suggests that they are likely to be responsible for the specific interactions with other proteins that determine in which photosystem and in which position within a photosystem complex the protein will be located (Schwartz et al. 1991).

In the N-terminal region, there is little overall conservation. The LHC II's and CP 29 Type I share a WYGPDR motif; LHC I's a WFPG concensus (boxes). The transit peptide cleavage sites, where determined for the same protein type in tomato or other species, are marked with arrow-heads (references in Schwartz et al. 1991). They do not appear to have the transit peptide cleavage motif deduced from a number of nuclear-encoded chloroplast proteins by Gavel and von Heijne (1990), except for the occurrence of a single R five or six residues N-terminal to the cleavage site.

From the beginning of the first conserved region to the end of the second (excluding the last 12 amino acids) the three LHC II Types are 78% identical and 87% similar. In this core region, CP 29 has 58% identity and 70% similarity with at least two of the LHC II's. In the first conserved region, the four LHC I Types are 35% identical (54% similar) to each other, but in the second conserved region they are 46% identical (61% similar). Because the connectors on either side of the second TMH are of different lengths, it is difficult to make a quantitative estimate of LHC I similarity in the middle region. In comparing each of these proteins to LHC II Type I, it can be calculated (Chothia and Lesk 1986) that the molecular cores will have the same overall fold within 0.8–1.2 Å rms deviation (Green et al. 1992a). This means that they will all have very similar three-dimensional structures and that most conclusions drawn from the pea Chl-protein electron diffraction studies will be valid for the other members of the CAB family.

It should be pointed out that the names used here for the tomato CAB proteins and their complexes are not the only ones in general use (cf. Bassi et al. 1990, Thornber et al. 1991). Fortunately, most of the groups publishing in the field have agreed on a standardized gene nomenclature, which is summarized in Table 1 (Jansson et al. 1992). That publication includes a useful table showing equivalent names for proteins and Chl-protein complexes in various systems, and another giving old and new names of all the published higher plant CAB genes.

Sequence comparisons: Higher plant and green algal CABs

In general, there is very little difference between proteins of a given Type from different higher plant species. A large compendium of angio-

Table 1. Green plant chlorophyll *a*/*b* proteins and genes

Role/Location	Complex	Polypeptide type	Gene name	Introns
Major antenna PS II	LHC II	Type I	*Lhcb1*	0
		Type II	*Lhcb2*	1
		Type III	*Lhcb3*	2
Core antenna PS II	CP 29(CP 26[a])	Type I	*Lhcb5*	5
		Type II	*Lhcb4*	1[b]
Minor antenna PS II	CP 24	–	*Lhcb6*	1
Antenna PS I	LHC I	Type I	*Lhca1*	3
		Type II	*Lhca2*	4
		Type III	*Lhca3*	2
		Type IV	*Lhca4*	2

[a] In some 'green gel' systems, the *Lhcb5* gene product is found in a 'CP 26' complex, and the *Lhcb4* is found in a redefined 'CP 29' complex (e.g. Bassi et al. 1990).
[b] Green and Pichersky (1993).

sperm LHC II Type I and II sequences was published several years ago (Chitnis and Thornber 1988). A few comparisons are available for other proteins: e.g. CP 29 Type I (*Lhcb5*) has been cloned from barley and tomato; CP 29 Type II (*Lhcb4*) from barley (Morishige and Thornber 1992) and *Arabidopsis* (Green and Pichersky 1993); LHC II Type III (*Lhcb3*) from barley, tomato and *Brassica*; CP 24 (*Lhcb6*) from tomato and spinach (refs. in Jansson et al. 1992, Boivin et al. 1993). A number of CAB genes have been cloned and sequenced from the gymnosperm *Pinus sylvestris* (Jansson and Gustafsson 1990, 1991). In comparing angiosperms and gymnosperms, the protein sequences of the same type were found to be >90% identical within the mature protein, with most of the substitutions in the N-terminal coding region. This implies that the different Types originated prior to the divergence of the angiosperms and gymnosperms (Jansson and Gustafsson 1991).

The sequences of a number of LHC II proteins cloned from lower plants and green algae are given in Fig. 2 along with the three tomato LHC II Types. A tripeptide motif at the end of the N-terminal region that appears to be distinctive for the three LHC II Types in higher plants (Jansson and Gustafsson 1991) is underlined. The fern (Pichersky et al. 1990) and moss (Long et al. 1989) sequences appear somewhat more like LHC II Type I than Type II but no definite conclusions should be drawn until more sequences become available. It has been suggested that one of the defective LHC II genes isolated from fern was derived from an LHC II Type III (Boivin et al. 1993).

Figure 2 also includes LHC II sequences from the green algae *Chlamydomonas* (Imbault et al. 1988, Larouche et al. 1991) and *Dunaliella* (Long et al. 1989, Laroche et al. 1990). The differences between the two *Dunaliella* sequences are not due to species differences but to the fact that different genes have been isolated (P. Falkowski, pers. comm.). It is impossible to assign them to any of the higher plant LHC II Types. However, their high degree of similarity to tomato LHC II sequences shows that they are definitely not CP 29 or LHC I sequences. A typical LHC II sequence from *Euglena gracilis* is just slightly more different from tomato than are the green algal sequences, with the exception of the short N-terminal region (Muchhal and Schwartzbach 1992). The euglenophytes are generally considered to be a separate line from the rest of the Chl *a*/*b*-containing green algae (chlorophytes) since they have three rather than two membranes surrounding the chloroplast (Gibbs 1978) and both LHC II and LHC I proteins are synthesized as polyprotein precursors that are cleaved to individual proteins after import into the chloroplast (Houlne and Schantz 1988, Muchhal and Schwartzbach 1992). However, the most striking thing about all the higher plant and algal LHC II sequences is their high degree of similarity, in length as well as in sequence. They are much more closely related to each other than tomato LHC II sequences are to tomato LHC I sequences (compare boxes in Figs. 1 and 2). This suggests that the LHC I–LHC II divergence occurred before the divergence of the lines leading to the modern chlorophytes, euglenophytes and land plants.

The fucoxanthin-Chl a/c *Proteins (FCPs)*

The eukaryotic algae are divided into three main groups (Christensen 1989) on the basis of their pigment composition: the Chl *b*-containing chlorophytes, the Chl *c*-containing chromophytes, and the Rhodophytes (red algae) which do not have Chl *b* or Chl *c* but use phycobilisomes for light-harvesting like the ancestral cyanobacteria. Before any chromophyte protein or gene sequences were available, immunological cross-reactivities between the fucoxanthin-Chl *a*/*c* (FCP) proteins of certain chromophytes and antisera raised to plant CAB proteins suggested that they had sequence similarities (Fawley et al. 1987, Caron et al. 1988, Hiller et al. 1988). The first complete gene sequences, obtained from the diatom *Phaeodactylum* (Grossman et al. 1990), showed clearly that FCPs were relatives of the CAB family (Fig. 2). With a combination of amino acid sequencing and PCR cloning, a family of at least four related but non-identical sequences have been obtained from the dinoflagellate *Amphidinium* (Hiller et al. 1993); one of them is shown in Fig. 2 and it is clearly related to the diatom sequence. In addition, several long peptide sequences have been obtained from the

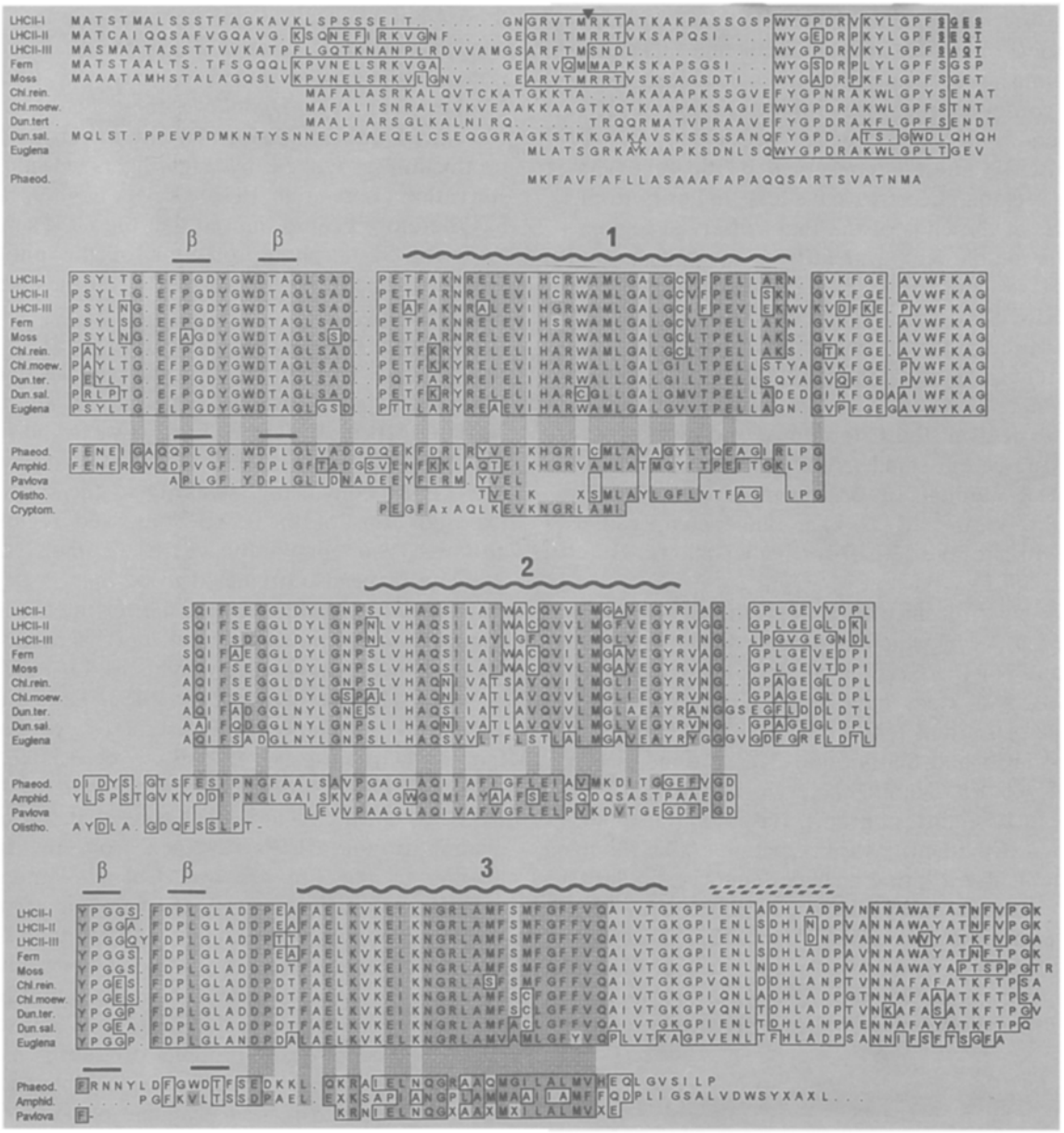

Fig. 2. Comparison of Chlorophyte LHCII and Chromophyte FCP Sequences. Boxed residues: identical or similar residues within Chl *a/b* or Chl *a/c* sequence comparisons; shaded residues: similar residues in both Chl *a/b* and Chl *a/c* sequences. Other symbols as in Fig. 1. LHC II sequences are those of tomato (Fig. 1); Fern, *Polystichum* **munitum** (Pichersky et al. 1990); Moss, *Physcomitrella patens* (Long et al. 1989); Chl. rein., *Chlamydomonas* ***reinhnrdtii*** (Imbault et al. 1988); Chl. moew., ***Chlamydomonas moewussii*** (Larouche et al. 1991); Dun. ter., *Dunaliella tertiolecta* (Laroche et al. 1990); Dun. sal. *Dunaliella salina* (Long et al. 1989); Euglena, ***E.*** *gracilis* (Muchhal and Schwartzbach 1992); Phaeod., ***Phaeodactylum tricornutum*** (Grossman et al. 1990); Amphid. ***Amphidinium*** *carterae* (Hiller et al. 1993); Pavlova, *P. lutherii* (Hiller et al. 1993); Olistho., *Olisthodiscus luteus* a.k.a. ***Heterosigma*** *carterae* (Green et al. 1992a); Cryptom., ***Cryptomonas*** sp. (Sidler et al. 1988).

prymnesiophyte (haptophyte) *Pavlova* (Hiller et al. 1993) and some shorter ones from the raphidophyte ***Olisthodiscus*** *luteus* ***(Heterosigma*** *carterae)* (Green et al. 1992a and unpublished results). The sequence from the cryptophyte ***Cryptomonus*** (Sidler et al. 1988) could be aligned in spite of its short length because it has the characteristic EVKNGRLAM motif.

In Fig. 2, shading is restricted to residues identical or similar in both CABs and FCPs. The relationship is strongest in the helical part of the two conserved regions, with the conserved H/N, R, and pair of E or E/Q residues. The F/WDPLGL motif is conserved in the first beta-turn region. The FCPs have less similarity to the CABs at the ends of the two conserved regions. In the middle region, the first Q/E of the second TMH is conserved, but the helix-terminating EXXR motif is not. The connectors linking the middle helix to the two conserved regions are shorter in the FCPs than the CABs. In addition, the *Phaeodactylum* FCPs are shorter by about 25 amino acids at the C-terminus, and do not have the predicted small amphipathic helix. FCPs from a number of organisms have molecular weights in the 18–22 kDa range compared to 25–28 kDa for LHC II's of higher plants (Hiller et al. 1991).

We have recently discovered that the red alga *Porphyridium*, which was not expected to have proteins of the CAB or FCP family since it has only Chl *a*, does have several polypeptides of 17–22 kDA that cross-react with both barley anti-LHC I and diatom anti-FCP antibodies (G. Wolfe, F.X. Cunningham, D. Durnford, B.R. Green, E. Gantt, in prep.). Based on our immunological evidence, we predict that the rhodophyte 'Chl *a*/*a*' proteins have sequence similarity to the CABs and FCPs even though they bind only Chl *a*. These findings imply that the ancestor of the CABs and FCPs was present prior to the divergence of the eukaryotic algae into rhodo-, chromo- and chlorophyte divisions.

The ELIPs

The Early-Light-Induced Proteins (ELIPs) were first discovered as the products of a class of mRNA turned on very rapidly after etiolated seedlings were first exposed to light (Meyer and Kloppstech 1984). In barley seedlings, ELIP mRNA levels peak 2–4 h after the start of illumination, i.e. before detectable amounts of CAB mRNAs are present (Grimm et al. 1989). The ELIP proteins are located in the thylakoid and have a half-life of less than 8 h. It has recently been discovered that ELIPs are not restricted to developing systems but are synthesized in response to high light stress in mature pea leaves (Adamska et al. 1992) and to dessication stress in Craterostigma (Bartels et al. 1992). In the green alga *Dunaliella* their synthesis is correlated with a massive increase in carotenoid production provoked by high light or sulfate starvation (Lers et al. 1991, Levy et al. 1992). It has therefore been proposed that the ELIPs may be involved in photoprotection of the photosynthetic apparatus, particularly during early development.

Like the CABs and FCPs, the ELIPs are predicted to have three TMH's (Fig. 3). They share the pair of conserved E's (or E/Q) and the NGRBAMB motif (where B = FILV) in the first and third TMH's (shaded residues). However, there is no convincing similarity to the stroma-exposed arms of the CAB conserved regions, although the comparable parts of the ELIP sequences are also predicted to be high in beta-turns. The ELIPs have much shorter interhelical connectors and no C-terminal 'tail', accounting for their lower molecular weights of 13–18 kDa in higher plants (Grimm et al. 1989). The ELIPs appear to be members of several related gene families (Grimm et al. 1989, Levy et al. 1992).

The ELIPs are not nearly as related to the CABs as the FCPs are, suggesting that the line leading to the ELIPs diverged from the line leading to the Chl *a*/*b* and Chl *a*/*c* antenna proteins before the emergence of the eukaryotic algal divisions. It is not known if ELIPs bind Chl or Chl precursors, as the proteins themselves have never been purified. (In the published studies, proteins were detected with antisera raised to recombinant proteins produced in *E. coli* or to synthetic peptides.) Since they have several conserved residues capable of being Chl ligands, they may be involved in Chl synthesis or in transferring Chl from its site of synthesis to the Chl proteins during thylakoid development. Alternatively, their connection with carotenoid biosynthesis in *Dunaliella* led to the suggestion they might be involved in the insertion of carotenoids into Chl-proteins (Lers et al. 1991).

*The 22 kDa protein of PS II (psb*S)

The 22 kDa intrinsic membrane protein of PS II (the product of the *psb*S gene) is always found as

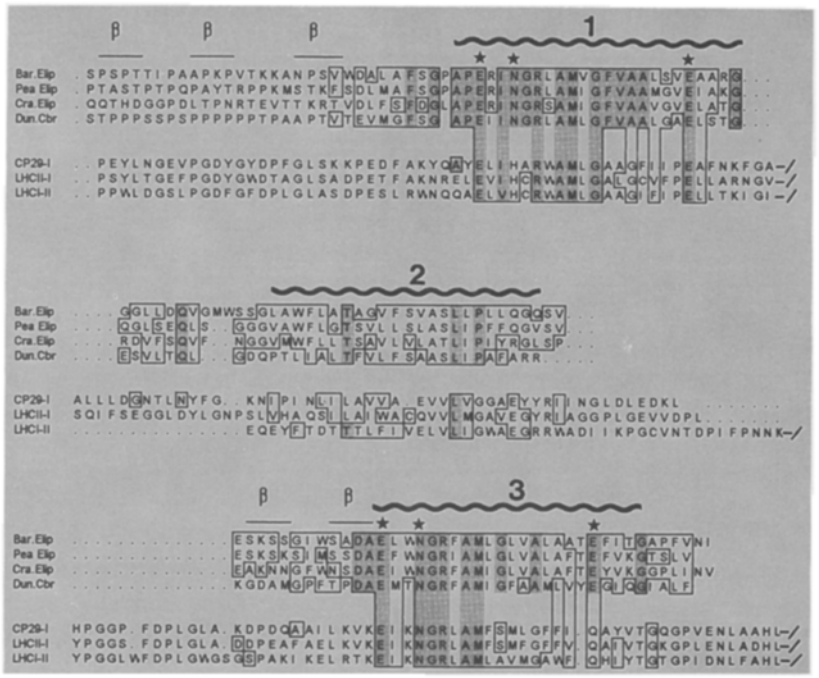

Fig. 3. Comparison of ELIPs with tomato CABs, starting with first conserved region. Bar., barley HV.58 (Grimm et al. 1989); Pea (Kolanus et al. 1987); Dun., *Dunaliella salina* carotenoid-biogenesis related protein (Lers et al. 1991); Cra, *Craferostigma dsp*-22 (Bartels et al. 1992). Boxed residues: identical (shaded) or similar residues shared by ELIPs or by ELIPs and CAB s. −/ part of CAB sequence omitted. Stars: possible Chl-ligating residues conserved in both ELIPs and CAB s. Other symbols as in Fig. 1.

part of PS II core particles but is not required for water-splitting activity (Ghanotakis et al. 1987). It was also detected immunologically in a cyanobacterium (Nilsson et al. 1990). When purified by ion-exchange chromatography it does not have any Chl associated with it (Ljungberg et al. 1986; Bowlby 1990). We were therefore surprised to find that antibodies raised to the Chl *a*/*b* complex CP 29 gave a weak positive reaction with the 22 kDa protein (Camm and Green 1989). Conversely, antibodies raised to ion-exchange purified 22 kDa protein cross-reacted weakly with both CP 29 polypeptides (Green and Camm 1990), suggesting that the apparent relatedness was not an artefact.

When the *psb*S gene of spinach was cloned and sequenced (Kim et al. 1992, Wedel et al. 1992) it was indeed found to be a member of the CAB family. However, the *psb*S protein has four TMHs rather than three! Figure 4 shows that the first three helices of the tomato *psb*S protein (M. Wallbraun, B.R. Green, B. Piechulla, E. Pichersky, in prep.) can be aligned with the three helices of tomato CAB proteins. The boxed residues are those that are similar to residues in any one of the CABs; shaded residues are identical. There are a number of shared residues in the first and third TMH's, including the R's and the pair of E's. However, the potentially Chl-ligating H and N are noticeable by their absence, as are the E's and terminal R in the second TMH (indicated by daggers in the figure). The exposed beta-turn domain in the first conserved region also appears to be missing, although it is possible that the last ten residues assigned to the N-terminal region, where they are aligned with CP 29, should actually be part of the beta-turn domain in the region below. There are a number of residues in the N-terminal region shared specifically by *psb*S and CP29 Type I, and in the middle region there are a few more matches between them than between *psb*S and any other CAB. However, since both of these proteins are N-terminally blocked, we do not know how many of these matches remain in the mature proteins.

The most striking thing about the *psb*S protein is the strong internal homology between the first

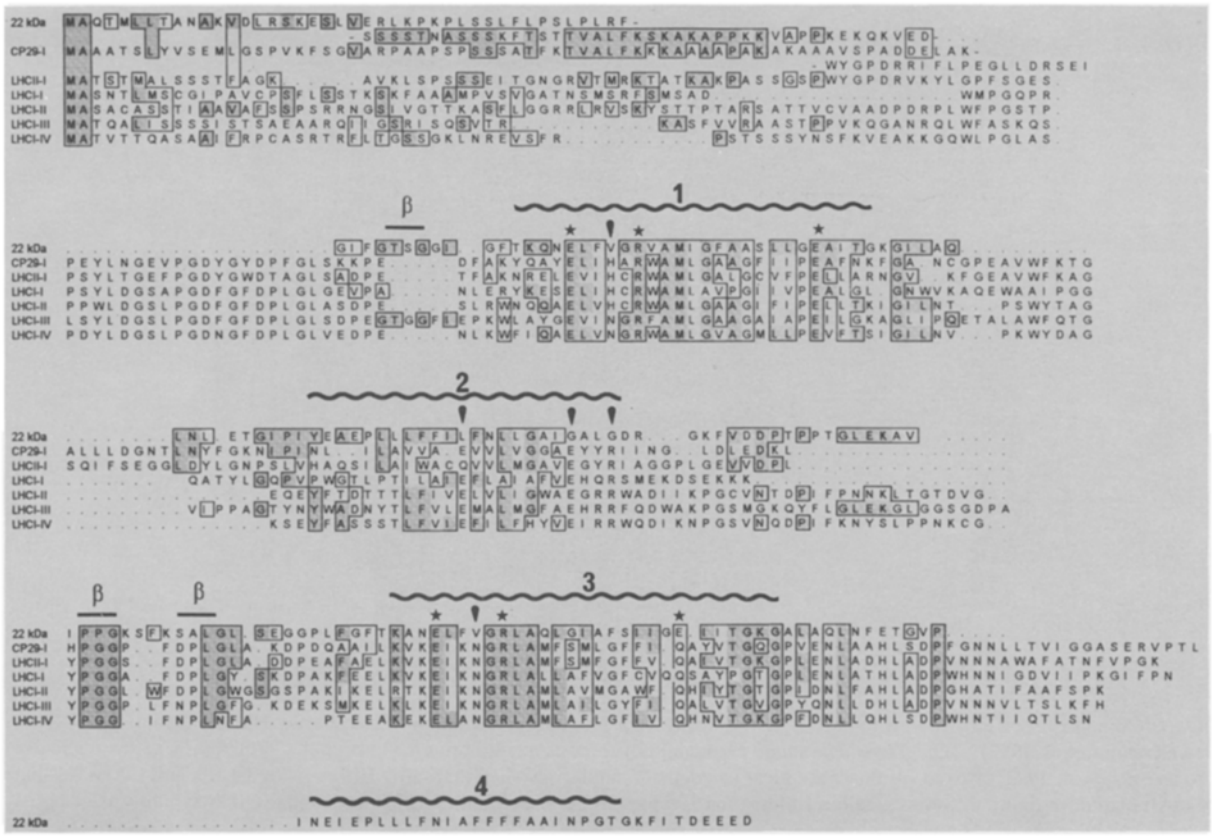

Fig. 4. Comparison of tomato *psb*S (22 kDa intrinsic protein of PS II) and CAB proteins. Shaded residues: identical in *psb*S and at least one CAB, boxed residues: similar. Dagger: significant residue not conserved in *psb*S. Other symbols as in Fig. 1.

two helices and the last two helices (Kim et al. 1992). The similarity between the first and third helices of the CABs was noted several years ago (Hoffman et al. 1987). However, the two halves of the 22 kDa protein are more related to each other than the comparable parts of the CABs are, and they have a higher degree of identity with each other than with the conserved regions of any of the CABs. Figure 5A compares the two conserved regions (first and third helices) in spinach *psb*S and in several tomato CABs. Figure 5B shows the high degree of similarity between the second and fourth TMH of *psb*S protein, and an attempt was made to align the C-terminal segment of CABs with their middle regions. There is a weak similarity in the case of CP 29 Type I, but it is not very convincing for the other two CABs. It is possible that the tail might represent a remnant of the connector between the first two helices.

Discussion

The internal duplication in *psb*S strongly suggests that the original four-helix ancestor arose from a two helix protein which underwent a tandem gene duplication (Fig. 6). The similarity between the first three helices of the four-helix *psb*S protein and the CABs suggests that they had a common four-helix ancestor. The lines leading to the three-helix CABs, FCPs, and ELIPs then lost the C-terminal helix by deletion. The original ancestor might not necessarily have been binding Chl, but some of its descendants could have been recruited into this role at a later time.

It is tempting to speculate that the ancestral two helix protein of this family could have originated by fusion of two single-helix proteins, similar to the alpha and beta subunits of the purple bacterial light-harvesting antennas. A

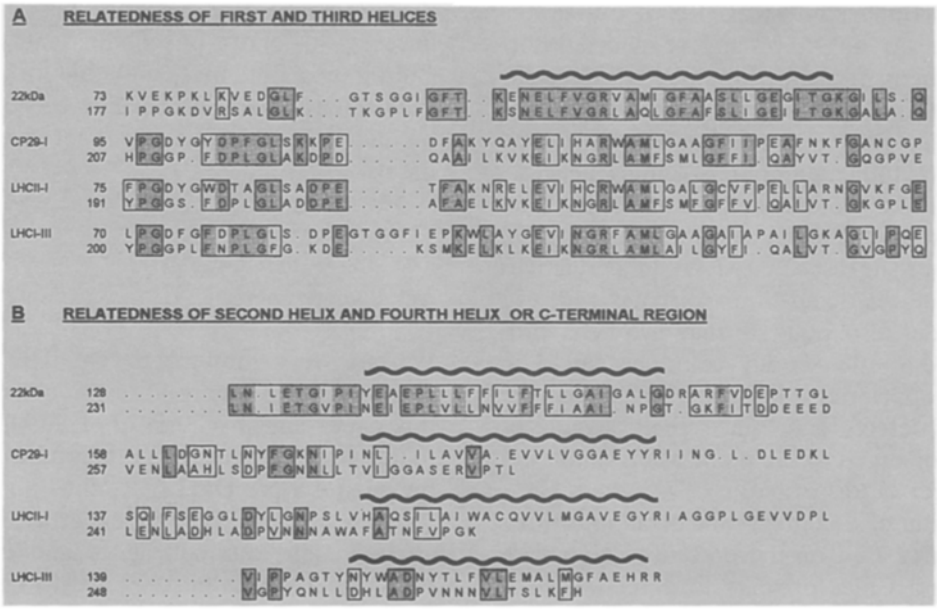

Fig. 5. Internal duplication in spinach *psb*S (22 kDa protein) and tomato CABs. Only partial sequences are shown. Boxed residues: identical or similar in both halves of the same protein. Numbers indicate number of first amino acid in precursor polypeptide. A: first and third helices and flanking sequences. B: second and fourth helices (*psb*S) or second helix and C-terminal sequence (CABs).

weak similarity between LHC I Type I and the bacterial beta subunit has been noted, but it is not found in the other CAB polypeptides (Hoffmann et al. 1987). Both alpha and beta polypeptides have their N-terminus on the same side of the membrane (Zuber and Brunisholz 1991), but it is possible to envisage mutational changes that would have made it possible to get a fusion protein to integrate properly into the membrane with its second helix effectively upside-down. It would be interesting to try to engineer such a protein in ***Rhodobacter sphaeroides.***

If it is confirmed that the 22 kDa protein is found in cyanobacteria, then the origin of the CAB/FCP/ELIP/*psb*S extended family may have preceded the symbiotic event(s) giving rise to the modern chloroplast. This also would mean that proteins related to CABs and FCPs could be found in cyanobacteria. In fact, there are several oxygenic prokaryotes, the prochlorophytes, that do have Chl ***b*** and do not have phycobilins (Lewin 1976). There is some discussion about whether they should be considered cyanobacteria that happen to have acquired Chl ***b*** and lost phycobilins or whether they are on a separate branch that is closer to the prokaryotic ancestor of the chloroplast than the other cyanobacteria (see Martin et al. 1992). A number of Chl ***a/b*** proteins of 34-36 kDa have been resolved in *Prochloron* (Hiller and Larkum 1985) and *Pro-*

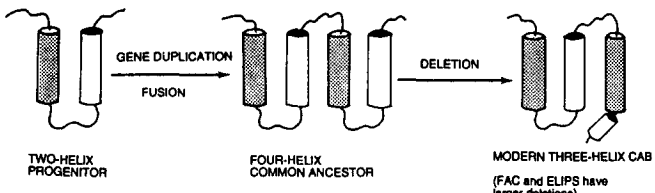

Fig. 6. Model for origin of CAB/ FCP/ ELIP family from a two-helix ancestor, via a common intermediate with four TMHs which also gave rise to the modern 22 kDa protein (*psb*S).

chlorothrix (Bullerjahn et al. 1987), but unfortunately we do not have any sequence information on them. In view of the higher molecular weight of these proteins compared to the CABs and FCPs, it will be interesting to see if they have three or four membrane-spanning helices.

The fact that proteins related to CABs and FCPs are present in a red alga that only has Chl *a* (see above) suggests that CAB/FCP relatives should not be restricted to prokaryotes with Chl *b*. Our model also suggests that two-helix proteins related to this family could be found in extant cyanobacteria, although it might be very difficult to detect them since their sequences could have diverged to the point where similarity can no longer be discerned.

The division of the eukaryotic algae into three major groups based on the presence or absence of Chl *b* and Chl *c* appears to agree with the majority of morphological, reproductive and biochemical characters (Christensen 1989). However, chlorophylls *a*, *b* and the two major variants of Chl *c* (Chl c_1 and Chl c_2) have only minor differences in ring substituents. The major difference between the Chl *c*'s and the other two Chls is that the former has no phytol tail and has an additional double bond in ring D. It is believed that all the chlorophylls are synthesized from the common precursors Mg 2,4-divinylpheoporphyrin (divinyl protochlorophyllide) and monovinyl protochlorophyllide (Jeffrey 1990). This suggests that it might not have been too difficult for members of a protein family to adapt to binding different types of chlorophyll.

In fact, every possible combination of light-harvesting pigments is found in at least one algal division (see Hiller et al. 1991, Jeffrey 1990). The eustigmatophytes have only Chl *a*; rhodophytes have Chl *a* and phycobilins. Chloro- and euglenophytes have Chl *a* and Chl *b*; the related micromonadophytes have Chl *a*, Chl *b* and Mg 2,4-divinylpheoporphyrin. All the Chl *c*-containing algae have Chl *a* and Chl c_2 but they may also have Chl c_1 and Chl c_3. Dinoflagellates have an additional antenna, the peridinen-Chl *a* protein complex. The cryptophytes have a Chl *a*/*c* antenna but also have phycobilins which are located in the lumen of the thylakoid and are not organized into phycobilisome structures, in contrast to rhodophytes and cyanobacteria. Assuming a monophyletic origin of the chloroplast in these lineages, this suggests that its prokaryotic ancestor had both phycobilins and a variety of Chl-binding antennas, and that its descendants subsequently lost the capacity to synthesize one or more of the possible light-harvesting pigments.

Acknowledgments

We are very grateful to Dr R.G. Hiller for allowing us to incorporate unpublished data in this paper and in Green et al. (1992a). We also thank Dion Durnford for helpful comments on the manuscript, Dr Tim Collins and Dr Stefan Jansson for sharing their expertise in computer analysis, and our colleagues and collaborators for sharing our enthusiasm about this interesting protein family. We thank the Natural Sciences and Engineering Council of Canada and the US Department of Agriculture for financial support.

References

Adamska I, Ohad I and Kloppstech K (1992) Synthesis of the early light-inducible protein is controlled by blue light and related to light stress. Proc Nat Acad Sci USA 89: 2610–2613

Bartels D, Hanke C, Schneider K, Michel D and Salamini F (1992) A desiccation-related Elip-like gene from the resurrection plant *Craterostigma plantagineum* is regulated by light and ABA. EMBO J 11: 2771–2778

Bassi R, Rigoni F and Giacometti GM (1990) Chlorophyll binding proteins with antenna function in higher plants and green algae. Photochem Photobiol 52: 1187–1206

Boivin R, Beauseigle D, Baszczynski CL and Bellemare G (1993) Isolation of *Lhcb3* sequences from *Brassica napus*: Evidence for conserved genes encoding LHC II Type III chlorophyll *a*/*b*-binding proteins. Genome 36: 139–146

Bowlby NR (1990) Alteration of secondary quinone acceptor activity in Photosystem II. PhD thesis, University of Michigan, Ann Arbor, MI, USA

Bullerjahn GS, Matthijs HCP, Mur LR and Sherman LA (1987) Chlorophyll-protein composition of the thylakoid membrane from *Prochlorothrix hollandica*: A chlorophyll *b*-containing prokaryote. Eur J Biochem 168: 295–300

Camm EL and Green BR (1989) The chlorophyll *ab* complex, CP 29, is associated with the Photosystem II reaction centre core. Biochim Biophys Acta 974: 180–184

Caron L, Remy R and Berkaloff C (1988) Polypeptide composition of light-harvesting complexes from some brown algae and diatoms. FEBS Lett 229: 11–15

Chitnis PR and Thornber JP (1988) The major light-harvest-

ing complex of Photosystem II: Aspects of its molecular and cell biology. Photosynth Res 16: 41–63

Chothia C and Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5: 823–826

Chou PY and Fasman GD (1974) Prediction of protein conformation. Biochemistry 13: 222–245

Christensen T (1989) The Chromophyta, past and present. In: Green J et al. (eds) The Chromophyte Algae: Problems and Perspectives, pp 1–12. Clarendon Press, Oxford

Coleman WJ and Youvan DC (1990) Spectroscopic analysis of genetically modified photosynthetic reaction centers. Annu Rev Biophys Biophys Chem 19: 333–367

Dunahay TG and Staehelin LA (1986) Isolation and characterization of a new minor chlorophyll *a*/*b*-protein complex (CP 24) of spinach. Plant Physiol 80: 429–434

Fawley MW, Morton SJ, Stewart KD and Mattox KR (1987) Evidence for a common evolutionary origin of light-harvesting chlorophyll *a*/*c*-protein complexes of *Pavlova gyrans* (Prymnesiophyceae) and *Phaeodactylum tricornutum* (Bacillariophyceae). J Phycol 23: 377–381

Gavel Y and von Heijne G (1990) A conserved cleavage-site motif in chloroplast transit peptides. FEBS Lett 261: 455–458

Ghanotakis DG, Waggoner CM, Bowlby NR, Demetriou DM, Babcock GT and Yocum CF (1987) Comparative structural and catalytic properties of oxygen-evolving Photosystem II preparations. Photosynth Res 14: 191–199

Gibbs SP (1978) The chloroplasts of *Euglena* may have evolved from symbiotic green algae. Can J Bot 56: 2883–2889

Green BR (1990) Structure prediction methods for membrane proteins: Comparison with the x-ray structure of *R. viridis* photosynthetic reaction centre. In: Villafranca JJ (ed) Current Research in Protein Chemistry, pp 395–404. Academic Press, New York

Green BR and Camm EL (1990) Relationship of Chl *a*/*b*-binding and related polypeptides in PS II core particles. In: Baltscheffsky M (ed) Current Research in Photosynthesis, Vol 1, pp 659–662. Kluwer Academic Publishers, Dordrecht, The Netherlands

Green BR and Pichersky E (1993) Nucleotide sequence of an *Arabidopsis thaliana* Lhcb4 gene. Plant Physiol 103: 1451–1452

Green BR, Pichersky E and Kloppstech K (1991) The chlorophyll *a*/*b*-binding light-harvesting antennas of green plants: The story of an extended gene family. Trends in Biochem Sci 16: 181–186

Green BR, Durnford D, Aebersold R and Pichersky E (1992a) Evolution of structure and function in the Chl *a*/*b* and Chl *a*/*c* antenna protein family. In: Murata N (ed) Research in Photosynthesis, Vol 1, pp 195–202. Kluwer Academic Publishers, Dordrecht, The Netherlands

Green BR, Shen D, Aebersold R and Pichersky E (1992b) Identification of the polypeptides of the major light-harvesting complex of Photosystem II (LHC II) with their genes in tomato. FEBS Lett 305: 18–22

Grimm B, Kruse E and Kloppstech K (1989) Transiently expressed early light-inducible thylakoid proteins share transmembrane domains with light-harvesting chlorophyll binding proteins. Plant Mol Biol 13: 583–593

Grossman AR, Manodori A and Snyder D (1990) Light-harvesting proteins of diatoms: The relationship to the chlorophyll *a*/*b* binding proteins of higher plants and their mode of transport into plastids. Mol Gen Genet 224: 91–100

Hiller RG and Larkum AWD (1985) The chlorophyll protein complexes of *Prochloron* sp. (Prochlorophyta). Biochim Biophys Acta 806: 107–115

Hiller RG, Larkum AWD and Wrench PM (1988) Chlorophyll proteins of the prymnesiophyte *Pavlova lutherii* (Droop) comb. nov.: Identification of the major light-harvesting complex. Biochim Biophys Acta 932: 223–231

Hiller RG, Anderson JM and Larkum AWD (1991) The chlorophyll-protein complexes of algae. In: Scheer H (ed) Chlorophylls, pp 529–547. CRC Press, Baton Rouge, LA, USA

Hiller RG, Wrench PM, Gooley AP, Shoebridge G and Breton J (1993) The major intrinsic light-harvesting protein of *Amphidinium*: Characterization and relation to other light-harvesting proteins. Photochem Photobiol 57: 125–131

Hoffman NE, Pichersky E, Malik VS, Castresana C, Ko K, Darr SC and Cashmore AR (1987) A cDNA clone encoding a Photosystem I protein with homology to Photosystem II chlorophyll *a*/*b*-binding polypeptides. Proc Natl Acad Sci USA 84: 8844–8848

Houlne G and Schantz R (1988) Characterization of cDNA sequences for LHC I apoproteins in *Euglena gracilis*: The mRNA encodes a large precursor containing several consecutive divergent polypeptides. Mol Gen Genet 213: 479–486

Imbault P, Wittemer C, Johanningmeier U, Jacobs JD and Howell SH (1988) Structure of the *Chlamydomonas reinhardtii* cabII-1 gene encoding a chlorophyll-*a*/*b*-binding protein. Gene 73: 397–407

Jansson S and Gustafsson P (1990) Type I and Type II genes for the chlorophyll *a*/*b*-binding protein in the gymnosperm *Pinus sylvestris* (Scots pine): cDNA clonging and sequence analysis. Plant Mol Biol 14: 287–296

Jansson S and Gustafsson P (1991) Evolutionary conservation of the chlorophyll *a*/*b*-binding proteins: cDNAs encoding Type I, II and III LHC I polypeptides from the gymnosperm Scots pine. Mol Gen Genet 229: 67–76

Jansson S, Pichersky E, Bassi R, Green BR, Ikeuchi M, Melis A, Simpson DJ, Spangfort M, Staehelin LA and Thornber JP (1992) A nomenclature for the genes encoding the chlorophyll *a*/*b*-binding proteins of higher plants. Plant Mol Biol Reporter 10: 242–253

Kim S, Sandusky P, Bowlby NR, Aebersold R, Green BR, Vlahakis S, Yocum CF and Pichersky E (1992) Characterization of a spinach *psb*S cDNA encoding the 22 kDa protein of Photosystem II. FEBS Lett 314: 67–71

Kolanus W, Scharnhorst C, Kuhne U and Herzfeld F (1987) The structure and light-dependent transient expression of a nuclear-encoded chloroplast protein gene from pea (*Pisum sativum* L.) Mol Gen Genet 209: 234–239

Kühlbrandt W and Wang DN (1991) Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. Nature 350: 130–134

Kühlbrandt W and Wang DN (1992) Photosynth Res 34: 81 (abstract)

Laroche J, Bennett J and Falkowski PG (1990) Characterization of a cDNA encoding for the 28.5 kDa LHC II apoprotein from the unicellular marine chlorophyte, *Dunaliella tertiolecta*. Gene 95: 165–171

Larouche L, Tremblay C, Simard C and Bellemare G (1991) Characterization of a cDNA encoding a PS II-association chlorophyll *a/b* binding protein (CAB) from *Chlamydomonas moewusii* fitting into neither type I nor type II. Curr Genet 19: 285–288

Lers A, Levy H and Zamir A (1991) Co-regulation of a gene homologous to early light-indiced genes in higher plants and beta-carotene biosynthesis in the alga *Dunaliella bardawail*. J Biol Chem 266: 13698–13705

Levy H, Gokhman I and Zamir A (1992) Regulation and light-harvesting complex II association of a *Dunaliella* protein homologous to early light-induced proteins in higher plants. J Biol Chem 267: 18831–18836

Lewin RL (1976) Prochlorophyta as a proposed new division of algae. Nature 261: 697–698

Ljungberg U, Akerlund H-E and Andersson B (1986) Isolation and characterization of the 10-kDa and 22-kDa polypeptides of higher plant Photosystem 2. Eur J Biochem 158: 477–482

Long Z, Wang S-Y and Nelson N (1989) Cloning and nucleotide sequence analysis of genes coding for the major chlorophyll-binding protein of the moss *Physcomitrella patens* and the halotolerant alga *Dunaliella salina*. Gene 76: 299–312

Martin W, Somerville CC and Loiseaux-de Goer S (1992) Molecular phylogenies of plastid origins and algal evolution. J Mol Evol 35: 385–404

Meyer G and Kloppstech K (1984) A rapidly light-induced chloroplast protein with a high turnover coded for by pea nuclear DNA. Eur J Biochem 138: 201–207

Morishige DT and Thornber JP (1992) Identification and analysis of a barley cDNA clone encoding the 31-kilodalton LHC IIa (CP 29) apoproteins of the light-harvesting antenna complex of Photosystem II. Plant Physiol 98: 238–245

Muchhal US and Schwartzbach SD (1992) Characterization of a *Euglena* gene encoding a polyprotein precursor to the light-harvesting chlorophyll-*a/b*-binding protein of Photosystem II. Plant Mol Biol 18: 287–299

Nilsson F, Andersson B and Jansson C (1990) Photosystem II characteristics of a constructed Synechocystis 6803 mutant lacking synthesis of the D1 polypeptide. Plant Mol Biol 14: 1051–1054

Pichersky E and Green BR (1990) The extended family of chlorophyll *a/b*-binding proteins of PS I and PS II. In: Baltscheffsky M (ed) Current Research in Photosynthesis, Vol 3, pp 553–556. Kluwer Academic Publishers, Dordrecht, The Netherlands

Pichersky E, Brock TG, Nguyen D, Hoffman NE, Piechulla B, Tanksley SD and Green BR (1989) A new member of the CAB gene family: Structure, expression and chromosomal location of CAB-8, the tomato gene encoding the Type III chlorophyll *a/b*-binding polypeptide of Photosystem I. Plant Mol Biol 12: 257–270

Pichersky E, Soltis D and Soltis P (1990) Defective chlorophyll *a/b*-binding protein genes in the genome of a homosporous fern. Proc Natl Acad Sci USA 87: 195–199

Pichersky E, Subramaniam R, White MJ, Reid J, Aebersold R and Green BR (1991) Chlorophyll *a/b* binding (CAB) polypeptides of CP 29, the internal chlorophyll *a/b* complex of PS II: Characterization of the tomato gene encoding the 26 kDA (type I) polypeptide, and evidence for a second CP 29 polypeptide. Mol Gen Genet 227: 277–284

Richardson JS and Richardson DC (1988) Amino acid preferences for specific locations at the ends of alpha-helices. Science 240: 1648–1652

Schiffer M and Edmundson AB (1967) Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. Biophys J 7: 121–135

Schuler GD, Altschul SF and Lipman DJ (1991) A workbench for multiple alignment construction and analysis. Proteins: Struct Funct Genet 9: 180–190

Schwartz E and Pichersky E (1990) Sequence of two tomato nuclear genes encoding chlorophyll *a/b*-binding proteins of CP 24, a PS II antenna component. Plant Mol Biol 15: 157–160

Schwartz E, Shen D, Aebersold R, McGrath JM, Pichersky E and Green BR (1991) Nucleotide sequence and chromosomal location of CAB11 and Cab12, the genes for the fourth polypeptide of the Photosystem I light-harvesting antenna (LHC I). FEBS Lett 280: 229–234

Sidler W, Frank G, Wehrmeyer W and Zuber H (1988) Structural studies on chlorophyll *a/c* light-harvesting complex from the cryptomonad *Cryptomonas maculata*: Partial amino acid sequences. Experientia 44: A60

Taylor WR (1986) The classification of amino acid conservation. J Theor Biol 119: 205–218

Thornber JP, Morishige DT, Anandan S and Peter GF (1991) Chlorophyll-carotenoid protein of higher plant thylakoids. In: Scheer H (ed) Chlorophylls, pp 549–585. CRC Press, Baton Rouge, LA, USA

Wedel N, Klein R, Ljungberg U, Andersson B and Herrmann RG (1992) The single-copy gene *psb*S codes for a phylogenetically intriguing 22 kDa polypeptide of Photosystem II. FEBS Lett 314: 61–66

Wilmot CM and Thornton JM (1988) Analysis and prediction of the different types of beta-turn in proteins. J Mol Biol 203: 221–232

Zuber H and Brunisholz RA (1991) Structure and function of antenna polypeptides and chlorophyll-protein complexes: Principles and variability. In: Scheer H (ed) Chlorophylls, pp 627–703. CRC Press, Baton Rouge, LA, USA