

Genome-Based Examination of Chlorophyll and Carotenoid Biosynthesis in *Chlamydomonas reinhardtii*^{1[w]}

Martin Lohr^{2*}, Chung-Soon Im², and Arthur R. Grossman

Institut für Allgemeine Botanik Johannes Gutenberg-Universität, 55099 Mainz, Germany (M.L.); and The Carnegie Institution Department of Plant Biology, Stanford, California 84305 (C.-S.I., A.R.G.)

The unicellular green alga *Chlamydomonas reinhardtii* is a particularly important model organism for the study of photosynthesis since this alga can grow heterotrophically, and mutants in photosynthesis are therefore conditional rather than lethal. The recently developed tools for genomic analyses of this organism have allowed us to identify most of the genes required for chlorophyll and carotenoid biosynthesis and to examine their phylogenetic relationships with homologous genes from vascular plants, other algae, and cyanobacteria. Comparative genome analyses revealed some intriguing features associated with pigment biosynthesis in *C. reinhardtii*; in some cases, there are additional conserved domains in the algal and plant but not the cyanobacterial proteins that may directly influence their activity, assembly, or regulation. For some steps in the chlorophyll biosynthetic pathway, we found multiple gene copies encoding putative isozymes. Phylogenetic studies, theoretical evaluation of gene expression through analysis of expressed sequence tag data and codon bias of each gene, enabled us to generate hypotheses concerning the function and regulation of the individual genes, and to propose targets for future research. We have also used quantitative polymerase chain reaction to examine the effect of low fluence light on the level of mRNA accumulation encoding key proteins of the biosynthetic pathways and examined differential expression of those genes encoding isozymes that function in the pathways. This work is directing us toward the exploration of the role of specific photoreceptors in the biosynthesis of pigments and the coordination of pigment biosynthesis with the synthesis of proteins of the photosynthetic apparatus.

Over the past several decades, the unicellular green alga *Chlamydomonas reinhardtii* has been an outstanding system for dissecting the function of various proteins involved in photosynthesis (Grossman, 2000; Harris, 2001; Rochaix, 2002). The ability of this alga to grow heterotrophically in the dark by metabolizing exogenous acetate has made it relatively easy to isolate a broad range of *C. reinhardtii* mutants that adversely affect photosynthetic function (Levine, 1969). Mutants defective for photosynthesis are readily analyzed at the genetic level as this organism has a relatively simple and short life cycle (Quarmby, 1994). Furthermore, a variety of physiological, biochemical, genetic, and molecular tools have been applied to studies of *C. reinhardtii*, making it an ideal model system for elucidating biological processes (for review, see Grossman, 2000; Harris, 2001; Rochaix, 2002; Grossman et al., 2004).

Recently, there has been considerable progress made with respect to genomic analysis of *C. reinhardtii*. The generation of extensive cDNA information (<http://www.chlamy.org/search.html>; Shrager et al., 2003), and a draft genome sequence (<http://genome.jgi-psf.org/chlre2>), is enabling researchers to understand more about genes present on the *C. reinhardtii* genome and the structure and expression of gene content. Furthermore, genome-based approaches have recently been applied to *C. reinhardtii* (<http://genome.jgi-psf.org/chlre2>) to elucidate the dynamics of the photosynthetic apparatus in response to nutrient and light conditions (Simpson and Stern, 2002; Grossman et al., 2003; Im et al., 2003; Shrager et al., 2003; Zhang et al., 2004; Y. Wang, Z. Sun, M.H. Horken, C.S. Im, Y. Xiang, A.R. Grossman, and D.P. Weeks, unpublished data).

Areas of interest with respect to light utilization in plants have focused on the involvement of pigments in both photosynthetic processes and the sensing and control of cellular processes through environmental light signals. Chlorophyll (Chl) and carotenoids are ubiquitous among photosynthetic organisms and play important roles in the function of the photosynthetic apparatus, the management of excitation energy and integration of photosynthetic function, and biogenesis of the photosynthetic membranes with the regulation of other cellular processes. Both Chl and carotenoid molecules bind to proteins integral to the photosynthetic machinery, where they absorb light energy to generate chemical bond energy (in the form of sugars) and also function in efficiently managing the use

¹ This work was supported by the Deutsche Forschungsgemeinschaft (grant no. LO840/1–1 to M.L.). A.R.G. would like to thank the National Science Foundation for supporting genomic research using *Chlamydomonas reinhardtii* (grant no. MCB 0235878). C.-S.I. was supported by the National Science Foundation (grant no. IBN 0084189 awarded to A.R.G.).

² These authors contributed equally to the paper.

* Corresponding author; e-mail lohr@uni-mainz.de; fax 49–6131–3923075.

[w] The online version of this article contains Web-only data.

Article, publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.104.056069.

of excitation energy. Carotenoids also participate in redox reactions (Tracewell et al., 2001; Frank and Brudvig, 2004), the protection of organisms from photodamage by quenching singlet oxygen and triplet Chl species (Siefermann-Harms, 1987; Frank and Cogdell, 1993; Yamamoto and Bassi, 1996; Formaggio et al., 2001; Baroli et al., 2003), and the dissipation of excess absorbed light energy via interactions with singlet excited Chl molecules (Demmig-Adams, 1990; Demmig-Adams et al., 1996; Yamamoto and Bassi, 1996; Niyogi, 1999; Baroli and Niyogi, 2000; Pogson and Rissler, 2000; Ma et al., 2003). Carotenoids may even help stabilize membrane structure (Havaux and Niyogi, 1999). Interestingly, intermediates in the Chl biosynthetic pathway may serve as signaling molecules that communicate the status of the pathway to the transcriptional machinery in the nucleus of the cell, thereby regulating levels of proteins that require Chl for their function (such as light-harvesting Chl-binding proteins; Johanningmeier and Howell, 1984; Johanningmeier, 1988; Kropat et al., 1997; Strand et al., 2003), and it appears that the biosynthesis of Chl is intimately linked to the presence and/or synthesis of

the light-harvesting complex (LHC) polypeptides (Xu et al., 2001). It is likely that Chl and carotenoid biosynthesis are precisely controlled to meet the demands of growing cells under a range of light conditions, and because intermediates in the former pathway are unstable and photoreactive, the accumulation of some intermediates in Chl biosynthesis can elicit the formation of damaging, reactive oxygen species. Although the synthesis of both Chl and carotenoids occurs within chloroplasts, in vascular plants all of the enzymes of the pathway are encoded by nuclear genes and are synthesized in the cytoplasm of the cell as precursor polypeptides with amino-terminal extensions (transit peptides) that enable them to pass through the double membrane of the chloroplast envelope and to their site of function within the organelle.

Chl is a cyclic tetrapyrrole coordinated by a central Mg²⁺ ion. The synthesis of Chl in plants and algae proceeds along the C₅ pathway, in which the first dedicated precursor of the pathway, 5-aminolevulinic acid (ALA), is synthesized from a Glu molecule (Fig. 1). Two molecules of ALA are then condensed to form porphobilinogen, and four porphobilinogen molecules are

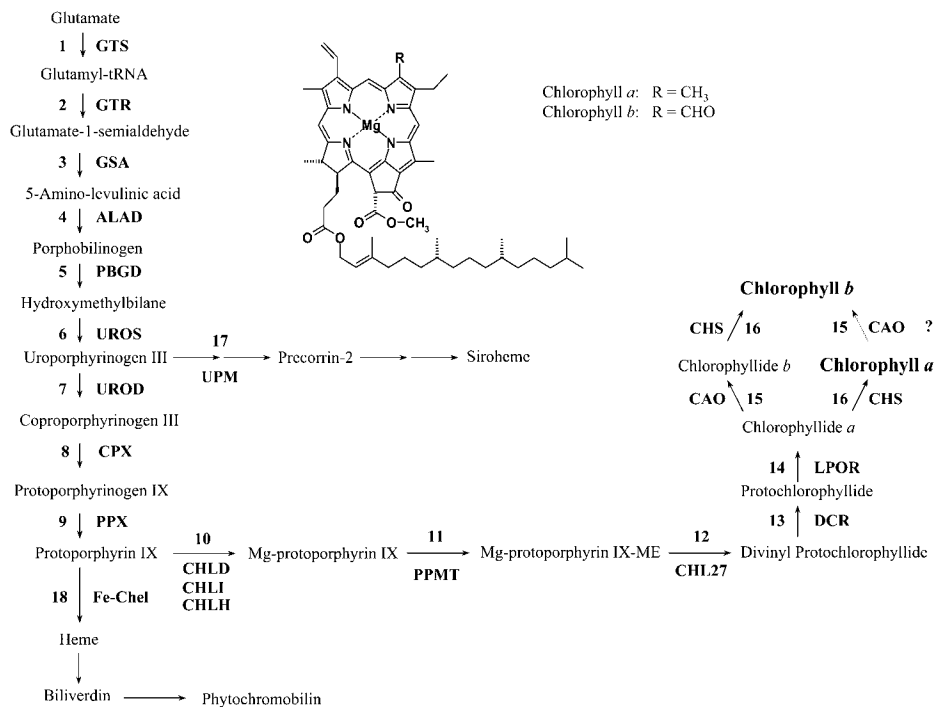


Figure 1. The Chl biosynthetic pathway. The names of the two major chlorophylls present in *C. reinhardtii* are in bold. Full enzyme names are as follows: GTS, glutamyl-tRNA synthetase; GTR, glutamyl-tRNA reductase; GSA, Glu-1-semialdehyde aminotransferase; ALAD, ALA dehydratase; PBGD, porphobilinogen deaminase; UROS, uroporphyrinogen III synthase; UROD, uroporphyrinogen III decarboxylase; CPX, coproporphyrinogen III oxidase; PPX, protoporphyrinogen IX oxidase; CHLD, protoporphyrin IX Mg-chelatase subunit D; CHLI, protoporphyrin IX Mg-chelatase subunit I; CHLH, protoporphyrin IX Mg-chelatase subunit H; PPMT, Mg-protoporphyrin IX methyltransferase; CHL27, Mg-protoporphyrin IX monomethylester cyclase subunit; DCR, divinyl protochlorophyllide reductase; LPOR, NADPH:protochlorophyllide oxidoreductase; CAO, chlorophyllide a oxygenase; CHS, Chl synthase; UPM, uroporphyrinogen III methyltransferase; and Fe-Chel, ferrochelatase. In the case of multiple commonly used synonyms, see Grossman et al. (2004); abbreviations indicative of enzymatic function were preferred. Throughout the text, the same abbreviations, in capital letters, are used for protein and gene designations, the latter being italicized. For chemical structures of intermediates in the pathway, see Beale (1999) or Vavilin and Vermaas (2002).

joined to form the first linear tetrapyrrole of the pathway, hydroxymethylbilane. The hydroxymethylbilane is then cyclized, followed by a decarboxylation and oxidation reactions to form protoporphyrin IX. Mg^{2+} is inserted into the protoporphyrin IX molecule, and the resulting Mg^{2+} protoporphyrin IX molecule is methylated, followed by a cyclization reaction that forms the cyclopentanone ring and sequential reduction steps to form chlorophyllide *a*. The reduction of protochlorophyllide to chlorophyllide can be catalyzed by two different enzymes, the nucleus-encoded, strictly light-dependent protochlorophyllide oxidoreductase (LPOR), common to all photosynthetic eukaryotes and cyanobacteria, or a light-independent (dark-active) enzyme complex (DPOR) that is not present in angiosperms. The latter is comprised of three subunits (ChlB, ChlL, and ChlN) that are encoded by the plastid genome. Phytylation of chlorophyllide *a* yields Chl *a*, while oxidation of chlorophyllide *a* could yield chlorophyllide *b* followed by phytylation to form Chl *b*. This pathway and its regulation have been reviewed re-

cently (Reinbothe et al., 1996; Suzuki et al., 1997; Beale, 1999; Vavilin and Vermaas, 2002; Cornah et al., 2003; Grossman et al., 2004).

The carotenoids are isoprenoids that belong to the tetraterpenoid group. Their basic structure is a C_{40} backbone containing a network of conjugated double bonds that form an extended π -electron system; this accounts for the ability of these molecules to absorb in both the UV and visible region of the light spectrum. Carotenoids that consist exclusively of hydrogen and carbon atoms are collectively termed carotenes. However, most naturally occurring carotenoids are oxygenated at one or more positions, placing them into the xanthophyll subgroup, which has been associated with managing the utilization of light energy in plants and algae (Demmig-Adams, 1990; Niyogi, 1999).

The biosynthesis of carotenoids (Fig. 2) starts with isopentenyl-diphosphate formation, the general precursor of all isoprenoids. In vascular plants and green algae, isopentenyl-diphosphate used for carotenogenesis is synthesized exclusively in the plastid by the

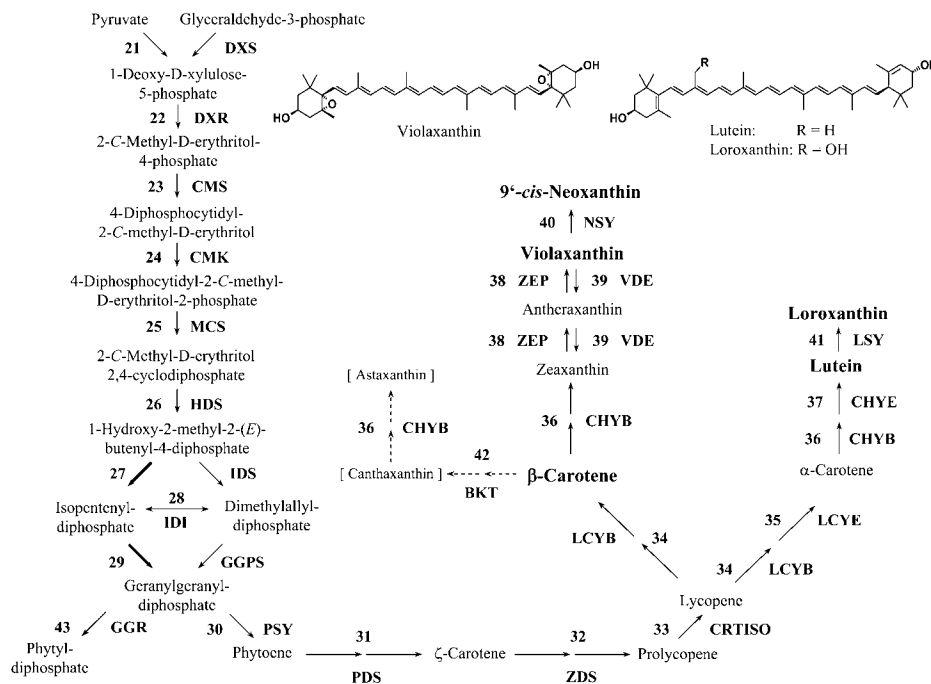


Figure 2. The carotenoid biosynthetic pathway. The names of the major carotenoids present in *C. reinhardtii* are in bold. Pigments in brackets have not yet been observed in *C. reinhardtii*. Full enzyme names are as follows: DXS, 1-deoxy-D-xylulose-5-phosphate synthase; DXR, 1-deoxy-D-xylulose-5-phosphate reductoisomerase; CMS, 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase; CMK, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; MCS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HDS, 1-hydroxy-2-methyl-2-(*E*)-butenyl-4-diphosphate synthase; IDS, isopentenyl diphosphate:dimethylallyl diphosphate synthase; IDI, isopentenyl diphosphate:dimethylallyl diphosphate isomerase; GGPS, geranylgeranyl pyrophosphate synthase; PSY, phytoene synthase; PDS, phytoene desaturase; ZDS, ζ -carotene desaturase; CRTISO, carotenoid isomerase; LCYB, lycopene β -cyclase; LCYE, lycopene ϵ -cyclase; CHYB, carotene β -hydroxylase; CHYE, carotene ϵ -hydroxylase; ZEP, zeaxanthin epoxidase; VDE, violaxanthin deepoxidase; NSY, neoxanthin synthase; LSY, loroxanthin synthase; BKT, carotene β -ketolase; and GGR, geranylgeranyl reductase. In the case of multiple commonly used synonyms, see Grossman et al. (2004); abbreviations indicative of enzymatic function were preferred. Throughout the text, the same abbreviations, in capital letters, are used for protein and gene designations, the latter being italicized. For chemical structures of intermediates in the MEP pathway, see Rodriguez-Concepcion and Boronat (2002), and for chemical structures of later intermediates, see Cunningham and Gantt (1998) or Hirschberg (2001).

recently discovered methylerythritol phosphate (MEP) pathway (Lichtenthaler, 1999; Rodriguez-Concepcion and Boronat, 2002; Rohmer, 2003). The first carotenoid, phytoene, results from the bonding of two C₂₀ molecules, each derived from the condensation of four C₅-isoprenoid units, to build the symmetrical C₄₀ backbone. This is followed by extension of the π -electron system through sequential desaturation steps and cyclization of the ends of the molecule to generate carotenes. Finally, the introduction of oxygen groups onto the molecule generates xanthophylls. Details of carotenoid biosynthesis have been the subject of several recent reviews (Cunningham and Gantt, 1998; Hirschberg, 2001; Grossman et al., 2004).

Whole-genome information is being generated for a number of photosynthetic eukaryotes (*Arabidopsis* Genome Initiative, 2001; Yu et al., 2002; Goff et al., 2002; Armbrust et al., 2004; Matsuzaki et al., 2004), and a nearly completed genome sequence (<http://genome.jgi-psf.org/chlre2>) as well as a wealth of cDNA information are available for *C. reinhardtii*. In this article, we exploit this genomic information to define the different genes encoding enzymes involved in Chl and carotenoid biosynthesis in *C. reinhardtii*, focusing on the relationship of the predicted protein sequences of this alga to those of *Arabidopsis* (*Arabidopsis thaliana*) and *Synechocystis* PCC 6803. The analyses are specifically restricted to nuclear genes that encode proteins with enzymatic activity in the biosynthetic pathways leading to the formation of Chl and carotenoids. We have learned about the structure of these genes and aspects of protein function based on comparisons of the deduced amino acid sequences, analyzed the encoded proteins for the presence of organellar targeting presequences, and identified different potential isozymes associated with specific reactions in the biosynthetic pathways. In addition, we have examined codon usage of the different genes, accumulation of the mRNAs derived from the different isogenes, and the influence of light on their expression levels. These analyses have enabled us to generate hypotheses concerning the function and regulation of proteins involved in the biosynthesis of both Chl and carotenoids.

RESULTS AND DISCUSSION

General Comparison of Chl and Carotenoid Biosynthetic Genes from *C. reinhardtii* with Similar Genes from *Arabidopsis* and *Synechocystis* PCC 6803

The genes predicted to encode most of the polypeptides known to be directly involved in the biosynthesis of Chl and carotenoids in vascular plants were identified in the current version (assembly v2.0) of the *C. reinhardtii* genome and GenBank expressed sequence tag (EST) entries as of August 2004. Features of these genes have been compiled and are summarized in Table I (Chl genes) and Table II (carotenoid genes). These tables are intended to provide readers with a summary of the information with respect to genes

encoding the enzymes of the Chl and carotenoid biosynthetic pathways and to serve as a resource to use for more in-depth analyses/experimentation. As indicated in the tables, some genes still contain gaps and/or have only partial cDNA coverage. Also, a number of gene models predicted from analysis of the genomic DNA sequence are incorrect, partly a consequence of sequence gaps, but also caused by non-canonical intron borders; often the correct mature transcript sequence can be inferred from available cDNA information. All gene models that we recognized as flawed are italicized in Tables I and II. Specific information on incorrect model prediction is included in the manual annotation of the respective gene models on the Joint Genome Institute (JGI) genome browser (<http://genome.jgi-psf.org/chlre2/chlre2.home.html>). Furthermore, we have performed additional cDNA sequencing for some of these genes to clarify or add needed sequence information (Tables I and II; see also "Materials and Methods").

Alignments of the predicted amino acid sequences from the homologous Chl and carotenoid biosynthesis genes from *C. reinhardtii*, the vascular plant *Arabidopsis*, and the cyanobacterium *Synechocystis* PCC 6803 were constructed and compared with respect to the lengths of the encoded proteins, their degree of conservation (expressed as percent identity/similarity), and the number of shared introns for the eukaryotic sequences. The presence of putative targeting presequences and additional conserved domains exclusively present in eukaryotic homologs were also investigated, with results of the analyses summarized in Tables I and II.

The predicted Chl and carotenoid biosynthesis genes from *C. reinhardtii* and *Arabidopsis* are consistently larger than those of *Synechocystis* PCC 6803, suggesting that the eukaryotic polypeptides may contain organellar-targeting presequences and/or additional domains within the mature proteins. This was further examined by aligning each of the predicted proteins from *C. reinhardtii* with homologous sequences from several vascular plants and cyanobacteria (alignments not shown); these alignments confirmed that the sequences from *C. reinhardtii* and vascular plants contain N-terminal extensions, usually between 30 and 90 amino acids, relative to the homologous cyanobacterial sequences. The sizes of the N-terminal extensions on the *C. reinhardtii* polypeptides are presented in Tables I and II. In some cases, an additional conserved N-terminal domain, probably part of the mature polypeptide, was present on the *C. reinhardtii* and *Arabidopsis* proteins, relative to the cyanobacterial homolog. This additional sequence probably evolved after the origin of plastids. For predicted proteins containing an additional conserved N-terminal domain that appears to be present in the mature protein, the presequence sizes specified in Tables I and II are marked with asterisks. The potential significance of these domains is discussed in more detail below.

In Tables I and II, presequence lengths, as inferred from the amino acid sequence alignments, are also

Table 1. Comparison of Chl biosynthetic genes of *C. reinhardtii*, the vascular plant *Arabidopsis*, and the cyanobacterium *Synechocystis* PCC 6803

C. reinhardtii genes encoding putative Chl biosynthetic enzymes (see Fig. 1 for full names of gene products) were analyzed for completeness and cDNA coverage, results being indicated by the following abbreviations: N, not available; P, partial; C, complete; C, complete cDNA sequences which were generated for this publication (see "Materials and Methods" for accession numbers); n.h., no homolog; n.i., not yet identified; and n.p., not present. Gene models and sequence lengths recognized as incomplete or erroneous are italicized, as are those data that are biased by this circumstance and therefore are preliminary; gene models and additional information can be found at <http://genome.jgi-psf.org/chlr2/chlr2.home.html> (use model number as search term under "Advanced Search"). For comparative analyses, homologous protein sequences of *C. reinhardtii*, *Arabidopsis*, and *Synechocystis* PCC 6803 were aligned with ClustalW, and the alignments corrected where necessary. Length of putative presequences was determined as N-terminal extension of *C. reinhardtii* proteins as compared to cyanobacterial homologs. Then, nonconserved ends at the N terminus (presequences) and C terminus of proteins were clipped to yield alignments containing only the putative functional cores. From these truncated alignments, the percentage of positions with identical (ident.) or identical plus similar (simil.) amino acid positions were calculated using the BioEdit software. Targeting prediction was done with the software tools TargetP (TarP), iPSORT (iPS), and Predotar (Pred), results being indicated by the following abbreviations: M, mitochondrial; P, plastid; and n, no targeting signal predicted.

Step	Gene Product	C.r. cDNA/Gene	C.r. Gene Model	(Orthol.) Genes from A.t.	Genes from 6803	Protein Length			A.t. %ident./simil.	Shared Introns (C.r./A.t.) ^a	6803 %ident./simil.	C.r. Presequences from		Target Prediction TarP/iPS/Pred
						C.r.	A.t.	6803				Alignments ^b	ChloroP ^c	
1	GTS	N/C	<i>C_440038</i>	At5g64050	sll0179	560	570	483	60/82	2 (15/12)	39/62	58	60	M/M/M
2	GTR	C/C	<i>C_420027</i>	At1g58290 At1g09940 At2g31250	slr1808	522	543 530 524	427	59/81 56/78 50/73	1 (1/2) 1 (1/2) 1 (1/2)	47/69	82*	42	M/M/n
3	GSA	C/C	<i>C_110168</i>	At5g63570 At3g48730	sll0017	463	474 472	411	70/85 70/84	0 (7/3) 0 (7/2)	73/87	52*	29	M/M/n
4	ALAD	C/P	<i>C_680014</i>	At1g69740 At1g44318	sll1994	390	430 406	327	61/81 54/75	1 (7/11) 1 (7/9)	44/68	58*	70	M/M/n
5	PBGD	C/C	<i>C_970063</i>	At5g08280	slr1887	349	382	320	69/82	1 (5/4)	42/67	39	34	P/M/M
6	UROS	C/P	<i>C_520061</i>	At2g26540	sll0166	297	321	538	44/70	2 (7/9)	19/49	See text	39	M/M/M
7	UROD1	C/C	<i>C_140105</i>	At2g40490 At3g14930	slr0536	390	394 418	350	60/82 51/74	2 (11/5) 1 (11/5)	59/78	47	28	M/M/M
	UROD2	P/C	<i>C_50056</i>	At2g40490 At3g14930	slr0536	394	394 418	350	56/77 62/81	0 (5/5) 1 (5/5)	56/76	48	53	P/M/M
	UROD3	C/C	<i>C_50202</i>	At2g40490 At3g14930	slr0536	398	394 418	350	47/72 47/69	0 (7/5) 0 (7/5)	54/76	41	35	P/M/n
8	CPX1	C/C	<i>C_390089</i>	At1g03475 At4g03205 ^d	sll1185	365	386 233 ^d	340	64/79 61/78 ^d	0 (5/6) 0 (5/3)	42/57	60	31	M/M/M
	CPX2	P/C	<i>C_680010</i>	At1g03475 At4g03205 ^d	sll1185	349	386 233 ^d	340	39/61 38/60 ^d	0 (6/6) 0 (6/3)	35/55	49	32	M/M/M
9	PPX	C/P	<i>C_330078</i>	At4g01690 At5g14220	n.h. (sll1237) ^e	563	506 508	n.h.	59/80 25/51	3 (14/8) 1 (14/17)	n.h.	+	49	P/P/P
10a	CHLD	P/C	<i>C_2200011</i>	At1g08520	slr1777	767	760	676	65/84	1 (11/14)	58/81	85	62	P/n/P
10b	CHLI1	C/C	<i>C_90171</i>	At4g18480 At5g45930	slr1030	417	424 418	357	75/87 73/86	0 (9/2) 0 (9/2)	78/90	75	54	P/M/M
	CHLI2	P/C	<i>C_490037</i>	At4g18480 At5g45930	slr1030	425	424 418	357	62/81 60/81	0 (8/2) 0 (8/2)	61/81	40	37	M/M/M
10c	CHLH1	C/C	<i>C_570057</i>	At5g13630	slr1055	1,399	1,381	1,331	67/86	0 (11/4)	66/85	72	36	P/M/M
	CHLH2	N/P	<i>C_1410006</i>	At5g13630	slr1055	?	1,381	1,331	?	?	?	?	?	?
11	PPMT	C/C	<i>C_1010073</i>	At4g25080	slr0525	326	312	230	62/80	0 (6/1)	54/74	95*	52	M/P/n
12	CHL27A	C/C	<i>C_490082</i>	At3g56940	sll1214 sll1874	407	409 358	358	63/80	0 (4/4)	57/78 49/72	46	35	M/M/M
	CHL27B	C/C	<i>C_100097</i>	At3g56940	sll1214 sll1874	407	409 358	358	68/83	0 (6/4)	59/79 49/72	46	34	M/M/M
13	DCR	n.i.	n.i.	n.i.	n.i.	?	?	?	?	?	?	?	?	?
14	LPOR	C/C	<i>C_280053</i>	At5g54190 At4g27440 At1g03630	slr0506	397	405 401 401	322	69/85 69/87 71/87	1 (2/4) 1 (2/3) 1 (2/4)	52/75	86	56	P/M/n
15	CAO	C/C	<i>C_1330043</i>	At1g44446	AB021313 ^f	645	536	276 ^f	60/80	0 (4/8)	52/74 ^f	320*	29	M/M/M
16	CHS	P/P	<i>C_190019</i>	At3g51820	slr0056	365	387	324	69/85	0 (5/13)	66/84	60	43	M/M/M

^aNumber of identical intron positions shared by the homologous proteins from *C. reinhardtii* (C.r.) and *Arabidopsis* (A.t.), followed by the total number of introns (in brackets) present in the protein from C.r. and A.t., respectively. Incomplete or preliminary data are italicized. ^bNumbers

denote length of N-terminal extension beyond conserved amino acid motifs in the alignments; *, presequence contains an additional conserved domain which probably is part of the mature protein (see text for further explanations). +, Protein contains putative presequence but no cyanobacterial homolog for comparison. ^cNumbers denote putative length of targeting presequence as predicted by the software tool ChloroP (see "Materials and Methods"). ^dIsogene coding for a significantly shorter protein due to early stop codon; for calculation of identity/similarity, also amino acids resulting from conceptual translation after the stop codon were considered. ^eGene code for putative PPX from *Synechocystis*

PCC 6803 as suggested in cyanobase; experimental proof still outstanding. ^fNo homolog in *Synechocystis* PCC 6803; instead, homologous protein from *P. hollandica* was used. ^gNo prediction possible due to missing N terminus.

Table II. Comparison of carotenoid biosynthetic genes of *C. reinhardtii*, the vascular plant *Arabidopsis*, and the cyanobacterium *Synechocystis* PCC 6803

Data for putative carotenoid biosynthetic genes (see Fig. 2 for full names of gene products) were compiled, analyzed, and presented as described for the Chl biosynthetic genes in the legend of Table I. (Note that abbreviations are the same as in Table I).

Step	Gene Product	C.r. cDNA/Gene	C.r. Gene Model	(Orthol.) Genes from A.t.	Genes from 6803	Protein Length			A.t. %ident./simil.	Shared Introns (C.r./A.t.) ^a	6803 %ident./simil.	C.r. Presequences from		Target Prediction TarP/iPS/Pred
						C.r.	A.t.	6803				Alignments ^b	ChloroP ^c	
21	DXS	C/C	<i>C_1950004</i>	At4g15560 At5g11380 At3g21500	sl11945	735	717	640	71/86 50/75 62/77	2 (12/8) 2 (12/9) 1 (12/8)	43/67	84	50	P/M/P
22	DXR	C/P	<i>C_70193</i>	At5g62790	sl10019	455	477	394	73/86	3 (8/11)	65/83	55*	29	P/M/M
23	CMS	P/C	<i>C_310123</i>	At2g02500	slr0951	319	302	230	58/79	0 (0/11)	38/59	86	65	P/M/M
24	CMK	P/C	<i>C_550077</i>	At2g26930	sl10711	347	383	315	55/75	2 (7/10)	27/53	49	39	M/M/M
25	MCS	C/C	<i>C_370022</i>	At1g63970	slr1542	207	231	161	76/90	2 (7/2)	43/73	47	30	M/M/M
26	HDS	C/P	<i>C_630052</i>	At5g60600	slr2136	681	741	403	64/80	2 (15/18)	29/41	16*	22	P/P/M
27	IDS	C/C	<i>C_1340041</i>	At4g34350	slr0348	465	466	406	54/75	1 (6/8)	51/75	42	34	M/M/M
28	IDI	C/P	<i>C_1540008</i>	At3g02780 At5g16440	n.h.	307	261	n.h.	43/62	0 (5/5)	n.h.	+	54	P/P/n
29	GGPS	P/C	<i>C_1150003</i>	At4g36810 (12)	slr0739	345	371	302	64/84	0 (7/0)	57/78	52	34	M/M/M
30	PSY	P/C	<i>C_140131</i>	At5g17230	slr1255	382	422	337	64/83	0 (4/5)	59/74	70*	34	M/M/M
31	PDS	C/C	<i>C_490019</i>	At4g14210	slr1254	564	566	472	72/85	3 (5/13)	67/85	73*	24	P/M/n
32	ZDS	C/P	<i>C_440086</i>	At3g04870	slr0940	582	558	489	65/80	3 (13/12)	63/79	79*	22	M/n/n
33	CRTISO	N/C	<i>C_1180035</i>	At1g06820	sl10033	591	595	501	61/78	2 (14/12)	59/78	65	40	M/M/n
34	LCYB	C/P	No model ^d	At3g10230	CAA52677 ^e	590	501	411 ^f	51/72	0 (10/0)	31/59 ^e	88*	39	M/M/P
35	LCYE	C/C	<i>C_270161</i>	At5g57030	n.p.	583	524	43/64	43/64	1 (11/10)	+	+	39	M/M/M
36	CHYB	C/C	<i>C_1280029</i>	At4g25700 At5g52570	AAA64983 ^f	297	310	176 ^f	53/76 51/76	2 (6/6) 2 (6/6)	26/42 ^f	130*	46	M/M/M
37	CHYE	N/P	<i>C_310063</i>	At3g53130	n.p.	594	566	?	32/56	?	?	+	39	M/M/M
38	ZEP	C/C	<i>C_50020</i>	At5g67030	n.p.	763	667	46/65	46/65	2 (9/15)	?	+	12	M/M/M
39	VDE	n.i.	n.i.	At1g08550	n.p.	?	462	?	?	?	?	?	?	?
40	NSY	n.i.	n.i.	n.i.	n.p.	?	?	?	?	?	?	?	?	?
41	LSY	n.i.	n.i.	n.p.	n.p.	?	?	?	?	?	?	?	?	?
42	BKT	C/P	<i>C_1280030</i>	n.p.	n.p.	444	444	444	444	444	444	74	24	P/P/M
43	GGR	C/C	<i>C_180159</i>	At1g74470	sl11091	504	467	407	73/87	0 (4/2)	66/83	79	81	P/M/n

^aNumber of identical intron positions shared by the homologous proteins from *C. reinhardtii* (C.r.) and *Arabidopsis* (A.t.), followed by the total number of introns (in brackets) present in the protein from C.r. and A.t., respectively. Incomplete or preliminary data are italicized. ^bNumbers denote length of N-terminal extension beyond conserved amino acid motifs in the alignments; *, presequence contains an additional conserved domain which probably is part of the mature protein (see text for further explanations). +, Protein contains putative presequence but no cyanobacterial homolog for comparison. ^cNumbers denote putative length of targeting presequence as predicted by the software tool ChloroP (see "Materials and Methods"). ^dNo gene model predicted, but partial *LCYB* gene sequences on scaffolds 235, 1434, and 263. ^eNo homolog in *Synechocystis* PCC 6803; instead, homologous protein from the cyanobacterium *Synechococcus* PCC 7942 was used. ^fNo homolog in *Synechocystis* PCC 6803; instead, homologous protein from the proteobacterium *Pantoea agglomerans* (a.k.a. *Erwinia uredovora*) was used.

compared to results from cleavage site prediction by ChloroP (Emanuelsson et al., 1999). In general, ChloroP predicts shorter presequences, and only in the case of glutamyl-tRNA synthetase (GTS), ALA dehydratase (ALAD), uroporphyrinogen III decarboxylase 2 (UROD2), 1-hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate synthase (HDS), and geranylgeranyl reductase (GGR) do the predicted cleavage sites fall within conserved regions of the aligned proteins. We are aware of only one enzyme, coproporphyrinogen III oxidase 1 (CPX1), for which the cleavage site has been determined directly by sequencing the N terminus of the mature protein (Quinn et al., 1999); in this case, the experimental data and ChloroP cleavage site prediction are congruent. However, in other cases, incorrect predictions by ChloroP are likely to occur, and only N-terminal sequencing of mature proteins will provide valid data on the lengths and cleavage sites for presequences.

We also analyzed the *C. reinhardtii* deduced protein sequences with the targeting prediction tools TargetP (Nielsen et al., 1997; Emanuelsson et al., 2000), ChloroP (Emanuelsson et al., 1999), Predotar (Small et al., 2004), and iPSORT (Bannai et al., 2002). Often these tools, developed primarily for use with vascular plant sequences, are not able to differentiate between *C. reinhardtii* mitochondrial and plastidic targeting signals since chloroplast transit peptides in this alga share features with both mitochondrial and plastid presequences of vascular plants (Franzen et al., 1990). However, in spite of the shortcomings of these programs, organellar targeting was predicted for nearly all of the proteins analyzed by at least two of the three software tools (Tables I and II). Carotenogenic ζ -carotene desaturase (ZDS; step 32; steps are marked in Figs. 1 and 2 and are listed in tables) was the sole protein for which only a single algorithm predicts its localization to an organelle.

The MEP-pathway enzyme HDS (step 26) appears to have an exceptionally short leader sequence, as deduced from both cDNA and genomic information. The open reading frame (ORF) contains an eight-amino acid sequence that precedes the first conserved motif (YCES). However, both TargetP and iPSORT predicted targeting of this polypeptide to the chloroplast, while Predotar suggested mitochondrial localization (Table II). Based on both TargetP and ChloroP, HDS has a putative organellar-targeting presequence of 22 amino acids, although the latter algorithm did not confirm that the presequence was involved in chloroplast localization. Since the conservation within the HDS polypeptide begins at amino acid nine and the transit peptide is predicted to be represented by the first 22 amino acids, it is conceivable that the HDS targeting sequence is not cleaved from the protein upon import into the chloroplast. This has recently been shown to be the case for CP29, a Chl-binding light-harvesting protein, of *C. reinhardtii* (Turkina et al., 2004).

Some Proteins Exhibit Significantly Less Conservation

Most proteins in Tables I and II are highly conserved among *C. reinhardtii*, Arabidopsis, and *Synechocystis* PCC 6803, sharing more than 60% pairwise amino acid identity and approximately 80% amino acid similarity. In many cases, the close phylogenetic relationship between *C. reinhardtii* and Arabidopsis genes is supported by the presence of one or more conserved intron positions. However, the deduced sequences for some Chl and carotenoid biosynthetic enzymes exhibit a significantly lower level of conservation. A lack of conservation is striking for the uroporphyrinogen III synthase (UROS; step 6); this protein is poorly conserved among all three of the organisms examined in this analysis.

In bacteria, UROS is the product of *hemD*. A number of (cyano)bacterial species, including *Synechocystis* PCC 6803, contain a *hemD*-like gene predicted to encode a hybrid protein representing a fusion of uroporphyrinogen III methyltransferase (UPM) with UROS (Panek and O'Brian, 2002). The occurrence of UROS as a domain of a fusion protein in *Synechocystis* PCC 6803 could in part explain the low similarity between the cyanobacterial UROS domain of the fusion protein and the UROS of *C. reinhardtii*. UPM catalyzes the first committed step in the biosynthesis of siroheme, which involves the methylation of the product formed by the UROS reaction (Fig. 1). Therefore, fusion of UPM with UROS in the (cyano)bacterial proteins probably has a role in regulating allocation of pathway intermediates for siroheme and Chl formation. Arabidopsis (At5g40850; Leustek et al., 1997) and *C. reinhardtii* (C_940006) both contain homologs of UPM; it is important to examine the expression of these proteins and their potential for interactions with UROS.

Several proteins that are part of the Chl biosynthetic pathway of *C. reinhardtii* are represented by multiple genes coding for putative isozymes. Some of these

predicted proteins, specifically the isoforms of UROD3 (step 7), CPX2 (step 8), and a putative H-subunit of the magnesium (Mg)-chelataze (CHLH2; step 10c), appear to have diverged significantly from their counterparts in Arabidopsis and *Synechocystis* PCC 6803. As a first approximation, this can be explained by a relaxed pressure to conserve genes that are represented by multiple copies on the genome, allowing for the evolution of enzymes with altered function(s) or expression patterns. A more detailed analysis of the potential isozymes is presented below.

With respect to the carotenoid biosynthetic pathway, the sequence of the enzyme isopentenyl diphosphate: dimethylallyl diphosphate isomerase (IDI; step 28) is not highly conserved. The *C. reinhardtii* and Arabidopsis enzymes are 43% identical and 62% similar at the amino acid sequence level. Low conservation for this protein was previously noted by Cunningham and Gantt (2000); they concluded that IDI from green algae and vascular plants were likely to have separate origins. The cyanobacterial (type II) IDI is completely unrelated to the eukaryotic enzyme (Steinbacher et al., 2003). Furthermore, the late carotenogenic enzymes in *C. reinhardtii* and Arabidopsis, including the lycopene cyclases (LCYB, step 34; LCYE, step 35), the carotenoid hydroxylases (CHYB, step 36; CHYE, step 37), and zeaxanthin epoxidase (ZEP; step 38), are significantly less conserved than the other enzymes in the pathway. These proteins have no cyanobacterial homologs, with the exception of the lycopene cyclases in the genus *Prochlorococcus*.

Several deduced proteins that function in Chl and carotenoid biosynthesis in *C. reinhardtii* and Arabidopsis have strong similarity to each other but low levels of conservation relative to their cyanobacterial homologs. The Chl genes in this group encode the plastidic GTS (step 1), the porphobilinogen deaminase (PBGD; step 5), and the CPX (step 8). The carotenoid genes in this group encode most enzymes of the MEP pathway (steps 21–27), namely, deoxy-xylulose-5-phosphate synthase (DXS), 4-diphosphocytidyl-2-methyl-erythritol synthase (CMS), 4-diphosphocytidyl-2-methyl-erythritol kinase (CMK), 2-methyl-erythritol-2,4-cyclodiphosphate synthase (MCS), and HDS. For all of these genes, identity/similarity between *Synechocystis* PCC 6803 and *C. reinhardtii* is about 20% lower than between *C. reinhardtii* and Arabidopsis (Tables I and II). The low similarity between MEP enzymes of vascular plants and cyanobacteria was noted previously (Lange et al., 2000); a similar degree of difference in these proteins between *C. reinhardtii* and *Synechocystis* PCC 6803 is noted here. Lange et al. (2000) suggested the most likely explanation for this observation to be a lateral transfer of these genes from other eubacteria into cyanobacteria subsequent to the primary endosymbiosis that led to evolution of extant plastids.

The nuclear genes encoding proteins involved in Chl and heme biosynthesis in *C. reinhardtii* may have originated either from an ancestral chloroplast or mitochondrion. Indeed, the CPX proteins from vascular plants and *C. reinhardtii* are more similar to human

and yeast CPX than to any of the cyanobacterial homologs. CrCPX1 has 55% (73%) and CrCPX2 has 43% (64%) identity (similarity) to CPX of human, and only 42% (57%) and 35% (55%) identity (similarity) with *Synechocystis* PCC 6803 CPX, respectively (see Table I and below). Similarly, the green algal and vascular plant PBGD are most closely related to the homologous enzyme from α -proteobacteria, which are considered to be among the closest known eubacterial relatives of mitochondria (Gray et al., 1999). The similarity between the plant enzyme and the mitochondrial PBGD from animals, however, is much lower, while a comparison of the latter with eubacterial sequences revealed highest similarity to the enzyme from cyanobacteria. To unravel this virtual paradox, more detailed studies will be necessary to elucidate the phylogenetic relationship of these enzymes from a larger set of taxa.

Conserved Domains in *C. reinhardtii* and Vascular Plant Proteins That Are Absent in the Cyanobacterial Homologs

Since both Chl and carotenoids are synthesized in plastids, N-terminal plastid targeting signals are expected to be associated with all nucleus-encoded proteins that participate in the synthesis of these pigments. The targeting signals generally display little or no conservation at the primary sequence level (von Heijne et al., 1989). Several enzymes involved in pigment biosynthesis in vascular plants were observed to have conserved domains absent from their cyanobacterial counterparts. These conserved domains are generally composed of 20 to 40 amino acids, and are mostly located at the N terminus of the protein between the targeting signal and the first common domain to the *Arabidopsis*/*C. reinhardtii* and cyanobacterial homologs. These additional sequences are part of the mature protein but are probably not essential for catalytic activity since they are absent in the cyanobacterial enzymes. Removal of the N-terminal extensions from phytoene synthase (PSY; Misawa et al., 1994), lycopene cyclases LCYB and LCYE (Hugueney et al., 1995; Cunningham et al., 1996), and the chlorophyllide *a* oxygenase (CAO; Nagata et al., 2004) did not result in a loss of enzymatic activity. Hence, these N-terminal extensions may have a regulatory function, either through interactions with metabolites or other polypeptides. This is supported to some extent by the finding that expression of tomato (*Lycopersicon esculentum*) PSY lacking the conserved N-terminal motif in *Escherichia coli* resulted in higher phytoene production than the expression of tomato PSY for which just the targeting sequence had been removed (Misawa et al., 1994).

In the following analyses, we focus on domains of Chl and carotenoid biosynthesis enzymes conserved between vascular plants and *C. reinhardtii* but not present in the bacterial homologs. Identification of domains conserved only within the green algal line-

age would require genomic/cDNA sequence information from additional green algal genera. In the Chl biosynthetic pathway, the three early enzymes glutamyl-tRNA reductase (GTR), Glu-1-semialdehyde aminotransferase (GSA), and ALAD (steps 2–4), as well as the Mg-protoporphyrin IX methyltransferase PPMT (step 11), possess an N-terminal conserved domain of 15 to 20 amino acids present in both *C. reinhardtii* and vascular plant enzymes (Supplemental Fig. 1). While GTR, ALAD, and PPMT probably acquired this sequence after the establishment of plastids within host cells, in the case of GSA the conserved domain is also present in some of the cyanobacterial homologs (e.g. species from the genus *Prochlorococcus*). In other cyanobacteria, including *Synechocystis* PCC 6803, remnant of the sequence still appears to be present contiguous to the N terminus of the ORF in the genome, but it appears to be no longer part of the ORF (see Supplemental Fig. 1).

In the carotenoid biosynthetic pathway, the enzymes 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR; step 22), phytoene desaturase (PDS; step 31), and ZDS (step 32) share conserved N-terminal extensions of up to 40 amino acids (Supplemental Fig. 2). Interestingly, the extensions associated with LCYB and LCYE display only a very low level of conservation between vascular plants and green algae but are well conserved within each of the two clades (data not shown). However, our alignments include sequences from only three green algae for LCYB (*C. reinhardtii*, *Volvox carteri*, and *Haematococcus pluvialis*) and two for LCYE (*C. reinhardtii* and *V. carteri*), with all of these sequences from the genus *Volvocales* (data not shown). Therefore, additional analyses of conserved domains associated with the two cyclases would benefit from a broader taxon sampling. As hypothesized by Grossman et al. (2004), the putatively conserved domains of LCYB and LCYE might interact with LHC apoproteins leading to altered enzyme activity (see also below). We suggest that the conserved domains on these polypeptides in vascular plants and algae represent targets for regulatory processes and exciting areas for future research.

CAO (step 15) from vascular plants contains a particularly large N-terminal extension. The first CAO gene sequenced was from *C. reinhardtii* (Tanaka et al., 1998); the identification of this sequence facilitated the subsequent identification of the homologous gene from *Arabidopsis* (Espineda et al., 1999; Rüdiger et al., 1999). Partial sequences of CAO genes from the prochlorophytes *Prochloron didemni* and *Prochlorothrix hollandica* were also obtained (Tomitani et al., 1999). Biochemical studies on *Arabidopsis* CAO revealed that, at least in vitro, only chlorophyllide *a* can be used as a substrate for catalyzing the formation of chlorophyllide *b* (Oster et al., 2000). A comparison of full-length CAO sequences from various organisms (Nagata et al., 2004) has demonstrated that the mature *Arabidopsis* and *Oryza sativa* enzymes contain an N-terminal extension with a highly conserved

A-domain of approximately 130 amino acids and a less conserved B-domain of 30 amino acids; both domains are absent from the prochlorophyte and *C. reinhardtii* CAO. Although this N-terminal extension was shown to be dispensable for the catalytic activity of the Arabidopsis enzyme, it was hypothesized to play a role in regulation of enzyme activity (Nagata et al., 2004). An interesting observation made by Vermaas and coworkers (Xu et al., 2001) was that the activity of CAO from Arabidopsis expressed in *Synechocystis* PCC 6803 could be strongly enhanced by coexpressing it with an apoprotein of the LHCII from pea (*Pisum sativum*). Although no stably assembled LHCII was detectable in the cyanobacterial transformants and the newly formed Chl *b* accumulated mainly in the core complexes of both photosystems, the results suggest that an interaction between CAO and apo-LHCII may modulate the activity of the enzyme (Xu et al., 2001). Nagata et al. (2004) have suggested that the A-domain of CAO is critical for this interaction to occur.

In the *C. reinhardtii* genome, a number of small EST sequences are located upstream and in close proximity to the CAO gene. A model predicted for CAO by GreenGenie (genie.8.14) suggests that the *C. reinhardtii* ORF, originally predicted by Tanaka et al. (1998), might extend into these small EST sequences, generating a coding region with an additional 600 bp at the 5' end of the gene and containing one additional intron. Another gene prediction tool, GENSCAN (<http://genes.mit.edu/GENSCAN.html>), also suggests the presence of an alternative start codon, extending the CAO N-terminal by 182 amino acids. We confirmed these predictions by sequencing a long *C. reinhardtii* cDNA clone (AV626430) and used the deduced sequence to search sequence reads from the *V. carteri* whole-genome shotgun (WGS) library. We were able to assemble the complete *V. carteri* CAO coding region from these reads. Figure 3 shows an alignment of the deduced CAO proteins from *C. reinhardtii*, *V. carteri*, *Dunaliella salina*, Arabidopsis, *O. sativa*, and *P. hollandica*; the *D. salina* sequence is likely to be incomplete. The N-terminal regions of CAO from green algae and vascular plants exhibit significant similarities. In addition, the secondary structures of the putative N-terminal extensions of CAO predict an extended α -helix that aligns between the algal and vascular plant proteins. Finally, examination of the presequence for a putative plastid targeting sequence also corroborates the presence of an N-terminal extension on the *C. reinhardtii* CAO. The original *C. reinhardtii* (BAA33964) and *D. salina* (BAA82481) CAO sequences deposited in GenBank were not predicted by TargetP, iPSORT, or Predotar to have a presequence that routes the protein to the plastid, while such a presequence was predicted by all three of these algorithms for the extended forms of CAO from *C. reinhardtii* and *V. carteri*. However, recent biochemical evidence from immunological analyses of the putative CrCAO polypeptide using an AtCAO antibody suggests that mature CAO from *C. reinhardtii* has an approximate molecular mass of

51 kD (Eggink et al., 2004). This would correspond to the theoretical mass of 51.4 kD calculated from the CAO sequence (BAA33964) determined by Tanaka et al. (1998). The predicted molecular mass of CrCAO with the conserved extension, after removal of the predicted 29-amino acid presequence (predicted by ChloroP), would result in a 69-kD mature protein. Therefore, it is critical to sequence the N-terminal part of the mature, chloroplast-localized protein.

If the additional conserved domain at the N terminus of the *C. reinhardtii* CAO is confirmed to be part of the mature polypeptide, it will be important to establish whether or not it interacts with the *C. reinhardtii* LHC polypeptides and the specificity of these interactions, if they occur. It is possible that the low degree of conservation between extensions of the green algal and vascular plant CAO can be explained by corresponding differences in the green algal and vascular plant LHC polypeptides (Elrad and Grossman, 2004), and a need for the two catalyzes to coevolve.

HDS (step 26), which catalyzes the penultimate step in formation of active isoprene by the MEP pathway, is another potential target for additional research that would help elucidate regulatory processes involved in pigment biosynthesis. The *C. reinhardtii* HDS gene model (Table II) predicts the presence of an extended insertion of about 260 amino acids with significant sequence similarity to the analogous domain from HDS of Arabidopsis (Querol et al., 2002). This domain is located in the central part of the protein. While this domain is absent from bacterial HDS, the Arabidopsis HDS can complement a HDS-null mutant of *E. coli* (Querol et al., 2002). The presence of this additional domain in the *C. reinhardtii* HDS was confirmed by sequencing a putative HDS cDNA (AV626792). An alignment of the HDS sequences from *C. reinhardtii*, Arabidopsis, and *Synechocystis* PCC 6803 is available as Supplemental Figure 3. The significance of the additional domain(s) in the plant and algal enzyme is not known, but it is interesting to note that the amino acid identity of this domain from *C. reinhardtii* and Arabidopsis is only 49%, while the remainders of the proteins are 74% identical (data not shown). It is reasonable to speculate that the inserted eukaryotic domain of HDS may be involved in regulation and that the details of this regulation may be somewhat different in *C. reinhardtii* and Arabidopsis.

Two Carotenogenic Genes of Vascular Plants Have No Identified Homologs in *C. reinhardtii*, while the Presence of Another Carotenogenic Gene on the *C. reinhardtii* Genome Was Unexpected

For most genes known to be directly involved in the biosynthesis of Chl and carotenoids in vascular plants, we were able to identify homologs in the current version of the *C. reinhardtii* genome. However, we were unable to identify *C. reinhardtii* genes encoding the plant enzymes violaxanthin deepoxidase (VDE; step 39) and neoxanthin synthase (NSY; step 40). Since the

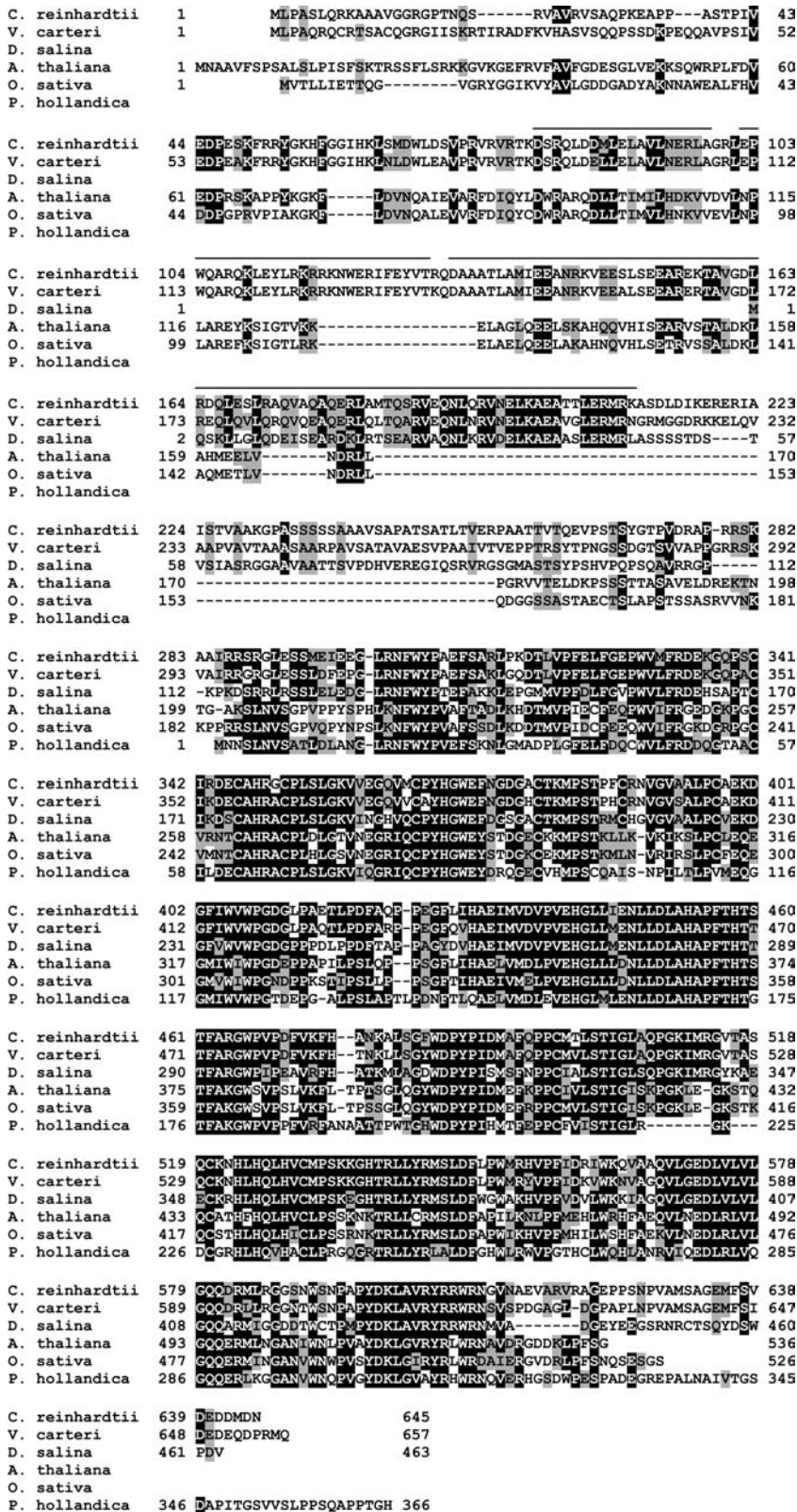


Figure 3. Protein alignment of chlorophyllide a oxygenase (CAO) from *C. reinhardtii* with the homologous enzymes from the green algae *V. carteri* and *D. salina* (sequence probably incomplete), the vascular plants *Arabidopsis* and *O. sativa*, and the full-length sequence of CAO from the prochlorophyte *P. hollandica*. At conserved sites, black boxes indicate identical amino acids, while gray boxes denote similar amino acids indicating conservative replacements (according to PAM250 matrix). Note the large N-terminal conserved domains of the putative full-length proteins from *C. reinhardtii* and *V. carteri* having significant sequence similarity to the corresponding domains in the CAO from the two vascular plants. Thick bars above these domains indicate putative α -helical stretches predicted by PSIPRED (McGuffin et al., 2000). The CAO cDNA sequence from *C. reinhardtii* has been deposited in GenBank under the accession number AY860816. Sequence accessions for other species are as follows: *V. carteri*, assembled from WGS reads ABSY196646.g1, AOBN193539.y1, AOBN79537.x1, AOBN-24721.x1, ABSY137769.b2, ABSY7818.g1, ABSY158303.x1, AOBN193539.x3, AOBN-24721.x1, ABSY203416.b1, AOBN191145.x1, AOBN43705.x1, ABSY3682.g2, ABSY23169.g1, AOBN163119.x1, AOBN14760.y1, AOBN-191145.y1; *D. salina*, BAA82481; *Arabidopsis*, BAA82484; *O. sativa*, AB021310; and *P. hollandica*, BAD02269.

current version of the *C. reinhardtii* genome is only about 90% complete, the missing genes might still be discovered in the fractions of the genome that have not yet been sequenced. However, we wouldn't regard this as very likely for reasons explained below.

NSY has only been identified in two species of the family Solanaceae, potato (*Solanum tuberosum*) and tomato (Al-Babili et al., 2000; Bouvier et al., 2000). In the complete genome of Arabidopsis, no gene homologous to NSY was detected (Hirschberg, 2001). Interestingly, NSY from tomato and potato turned out to be paralogous to the two lycopene cyclases (LCYB and LCYE), common to all vascular plants, and the closely related capsanthin-capsorubin synthase (CCS) from bellpepper (*Capsicum annuum*). Furthermore, both NSY (Ronen et al., 2000) and CCS (Hugueney et al., 1995) were shown to possess lycopene-cyclase activity. Therefore, it is conceivable that in plants lacking a separate NSY, one of the two lycopene cyclases (most likely LCYB based on its similarity to NSY) might be responsible for the formation of neoxanthin from violaxanthin, possibly triggered by interactions with the neoxanthin-binding proteins of LHCII (Grossman et al., 2004). Alternatively, an enzyme unrelated to NSY of the Solanaceae might be responsible for neoxanthin formation in Arabidopsis and *C. reinhardtii*.

VDE catalyzes the deepoxidation of violaxanthin as part of the photoprotective xanthophyll cycle (Yamamoto et al., 1999). Genes encoding VDE have been sequenced from several vascular plants (Bugos et al., 1998). Neither the current version of the *C. reinhardtii* genome nor WGS reads available from the closely related alga *V. carteri* contain any sequences with significant similarity to VDE from vascular plants. These results suggest that green algae may use a deepoxidating enzyme with characteristics different from those of the plant enzyme. This suggestion is supported by the observation that dithiothreitol, a potent inhibitor of vascular plant VDE, does not prevent violaxanthin deepoxidation in high light-exposed cultures of *C. reinhardtii* (K. Niyogi, personal communication). Identification of the deepoxidase from *C. reinhardtii* by map-based cloning is currently in progress (Anwaruzzaman et al., 2004).

We surprisingly detected a gene coding for a putative β -carotene ketolase (BKT; step 42), based on similarity to BKT from the green alga *H. pluvialis*. BKT introduces a keto-group at C(4) of β -ionon rings and, in conjunction with β -carotene hydroxylase (CHYB), catalyzes the formation of the ketocarotenoid astaxanthin (Lotan and Hirschberg, 1995; Breitenbach et al., 1996). Interestingly, the genes coding for CHYB and BKT are contiguous on the *C. reinhardtii* genome, with BKT located on the same strand and just upstream of CHYB. To the best of our knowledge, astaxanthin has not been detected in *C. reinhardtii*, and our attempts to detect it (using HPLC) in both nutrient-replete and nutrient-limited cultures have been unsuccessful (data not shown). The putative

BKT gene in *C. reinhardtii* appears to be expressed since it is represented by a cDNA clone (1024014H04) in the EST database. Since the BKT gene in the current version of the *C. reinhardtii* genome database contains two large gaps, we sequenced the corresponding cDNA. As the alignment in Figure 4 demonstrates, the central part of the *C. reinhardtii* and *H. pluvialis* homologs is highly conserved at the amino acid level (70% identity and 82% similarity). However, BKT from *C. reinhardtii* is predicted to have a C-terminal extension of about 115 amino acids, which is absent from any ketolase previously characterized. It will be interesting to examine the functional significance of this amino acid extension, which might relate to the absence of astaxanthin in *C. reinhardtii*.

***C. reinhardtii* and Arabidopsis Differ Significantly in the Number of Putative Isozymes Involved in the Biosynthesis of Chl and Carotenoids**

In Arabidopsis, there are often multiple genes coding for putative isozymes that function at a number of different steps in the pathway for Chl synthesis (Lange and Ghassemian, 2003). By contrast, most reactions in the analogous pathway in *C. reinhardtii* are catalyzed by unique gene products, with the exception of UROD, CPX, CHLI, CHLH, and CHL27 (Table I). The reactions of the carotenogenic pathway in *C. reinhardtii* are all catalyzed by unique gene products (with the caveat that there are still some gaps in the genome sequence). In Arabidopsis, there do appear to be isogenes for DXS, IDI, geranylgeranyl diphosphate synthase (GGPS; step 29), and CHYB (Table II).

The increased number of isozymes associated with pigment biosynthesis in vascular plants relative to *C. reinhardtii* or cyanobacteria may be related to increased regulatory demands and perhaps also to different local environments (e.g. in cells of different tissue types). As an example for the Chl biosynthesis pathway, the expression of *GTR1* (*HEMA1*) in Arabidopsis was highest in green tissue and under stringent light control, while *GTR2* was mainly expressed in roots and flowers in a light-independent manner (McCormac et al., 2001; Ujwal et al., 2002). Organ-specific expression of two GTR isogenes from barley (*Hordeum vulgare*; Bougri and Grimm, 1996) has also been reported. With respect to the carotenoid pathway, DXS was shown to be encoded by two different genes in the legume *Medicago truncatula*. *DXS1* was expressed in a variety of developing tissues, with the exception of the roots. *DXS2* expression was strongly stimulated in roots upon colonization with mycorrhizal fungi (Walter et al., 2002). There are both organ- and organellar-specific isoforms of GGPS in Arabidopsis (Zhu et al., 1997; Okada et al., 2000). There appear to be 12 different GGPS isogenes in Arabidopsis (Lange and Ghassemian, 2003), although expression of only five has been demonstrated (Okada et al., 2000). There is one GGPS and three other ORFs on the *C. reinhardtii* genome that encode related, prenyl transferase-like proteins. Alignment of these

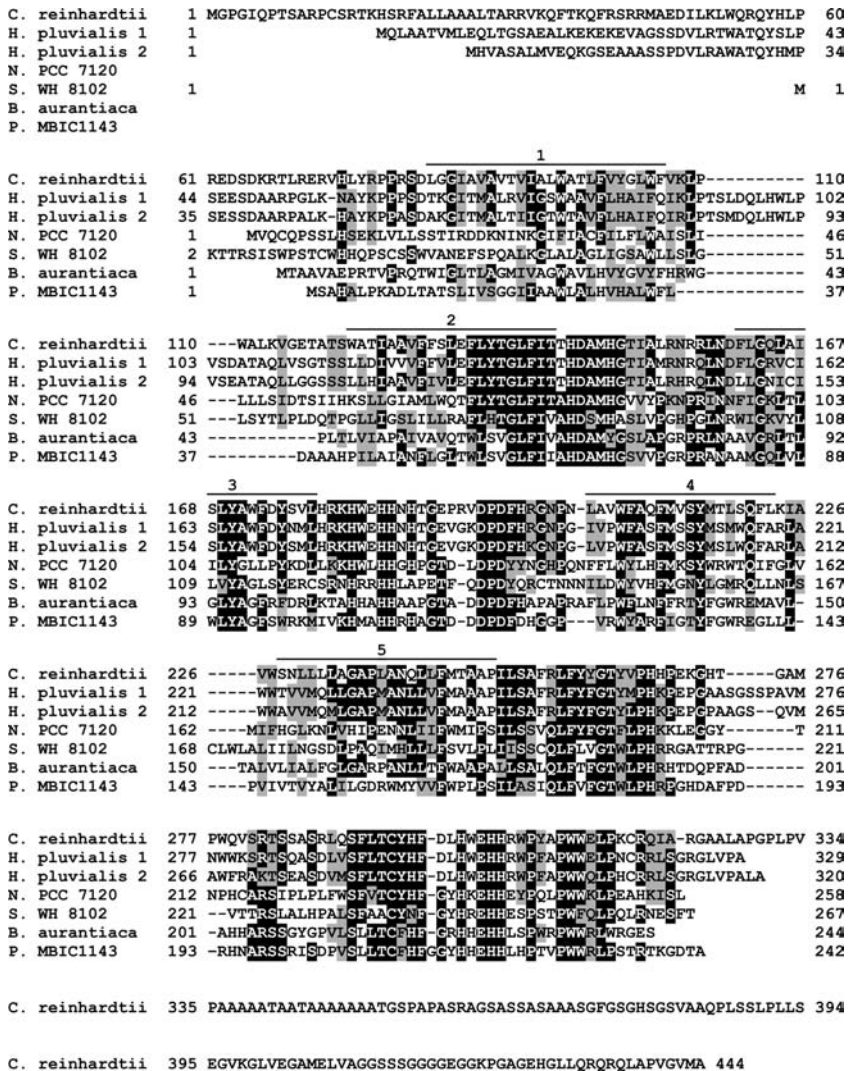


Figure 4. Protein alignment of BKT from *C. reinhardtii* with the homologous enzymes from the green alga *H. pluvialis* and the ketolases (CrT) from cyanobacteria and proteobacteria. At conserved sites, black boxes indicate identical amino acids, while gray boxes denote similar amino acids (conservative exchanges according to PAM250 matrix). Putative α -helical transmembrane domains predicted by TMHMM in the BKT from *C. reinhardtii* are indicated by bars above the sequence (see also Cunningham and Gantt, 1998). Note the large C-terminal extension of the protein from *C. reinhardtii* not present in the other proteins. The cDNA sequence from *C. reinhardtii* coding for BKT has been deposited in GenBank under the accession number AY860820. Other species and sequence accessions are as follows: *H. pluvialis* 1, CAA60478, and 2, BAA08300; *Nostoc PCC 7120*, BAB74888; *Synechococcus WH 8102*, ZP_00115639; *Brevundimonas aurantiaca*, AAN86030; and *Paracoccus MBIC1143*, BAA09591.

sequences with GGPS and related prenyl transferases from Arabidopsis and other vascular plants (data not shown) has tentatively enabled us to assign the products of the other genes the functions geranyl pyrophosphate synthase (gene model C_490103), farnesyl pyrophosphate synthase (C_120115), and solanesyl pyrophosphate synthase (C_1690011).

Putative Isozymes That Function in Chl Biosynthesis in *C. reinhardtii* and Their Relationship to Enzymes from Other Photosynthetic Organisms

The *C. reinhardtii* isogenes involved in Chl biosynthesis are *UROD* (step 7), *CPX* (step 8), two subunit genes of the Mg-chelatase (step 10), *CHLI* and *CHLH*, and the recently identified *CHL27* (step 12). The *CHL27* protein appears to be involved in catalyzing the formation of the cyclopentanone ring of Chl (Moseley et al., 2000; Tottey et al., 2003). As a first step toward a detailed characterization of potential isozymes associated with pigment biosynthesis in *C.*

reinhardtii, we searched for homologs in the genomes of Arabidopsis and *O. sativa*, in vascular plant EST databases, in the genomes of the red alga *Cyanidioschyzon merolae*, the diatom *Thalassiosira pseudonana*, as well as in the current cyanobacterial databases.

In *C. reinhardtii*, *UROD* is the first enzyme in the Chl biosynthetic pathway encoded by multiple genes (step 7), three in this case. A comparison among the predicted *UROD* proteins of *C. reinhardtii* is presented in Supplemental Figure 4. The encoded proteins have 43% to 55% identity and 67% to 76% similarity among themselves, and expression of all three of the *UROD* isogenes is supported by EST sequence data. Vascular plants in general appear to contain at least two different *UROD* genes; two isogenes were identified in Arabidopsis, potato, tobacco (*Nicotiana tabacum*), and barley. For *O. sativa* and *Zea mays*, cDNA data suggest the occurrence of three isogenes (but see below). The genome of *C. merolae* also contains two *UROD* isogenes, while the *T. pseudonana* genome harbors three isogenes. The cyanobacterial genomes (eight

complete and five partial) each contain a single *UROD* gene.

The phylogenetic relationship among *UROD* isoforms from different organisms is depicted in Figure 5. The cyanobacterial enzymes cluster at the base of the neighbor-joining tree, while the eukaryotic enzymes fall into three groups, with the CrUROD1 from *C. reinhardtii* located at the base of a cluster also containing *UROD1* from vascular plants. Similarly, CrUROD2 from *C. reinhardtii* and vascular plants form a second cluster. Both clusters have high bootstrap support. The third *UROD* cluster is divided into two subclusters containing the red algal and diatom isozymes, and while the phylogenetic position of CrUROD3 from *C. reinhardtii* is less well resolved, it appears to fall into a subcluster with one each of the red algal and the diatom isoforms. The other *UROD* from *C. merolae* and the two remaining isoenzymes from *T. pseudonana* (TpUROD3 was assembled from unplaced WGS reads) comprise the other subcluster. A very similar branching pattern, with similar bootstrap values, resulted from a maximum-likelihood analysis employing 100 bootstrap replicates (data not shown).

Surprisingly, OsUROD3 (from *O. sativa*) is most closely related to CrUROD2. However, the OsUROD3 sequence is supported by a single cDNA entry (AK110601), and we failed to retrieve any additional EST or genomic sequence data for this putative gene from the *O. sativa* databases. Hence, the single cDNA may represent a contamination of the cDNA library with an unidentified green alga. This is corroborated by comparative analyses of GC content and codon usage of the *O. sativa* *UROD* genes. While OsUROD1 and OsUROD2 have a GC content of 50% and 54% and an effective number of codons (ENC) used of 55.6 and 57.9, respectively, the OsUROD3 sequence has a strong bias both with respect to GC (64%) and ENC (36.2) values, which are more similar to the values expected for ORFs of green algae like *C. reinhardtii* or *V. carteri* (see Table III and below).

The *UROD* reaction is positioned at a branch point of tetrapyrrole biosynthesis, competing with UMP for the substrate uroporphyrinogen III. A comparison of intron positions among the *C. reinhardtii* and Arabidopsis isoforms (Table I; Supplemental Fig. 4) reveals that some intron positions are conserved between the different *UROD* isoforms within a given organism. These findings suggest that a gene duplication occurred after the endosymbiotic event that presaged the evolution of the chloroplast in eukaryotic plant cells. Furthermore, since all plant and algal species that we examined contain at least two different genes encoding putative *UROD* isozymes, it is possible that the *UROD* isozymes fulfill different roles in the cell; they may no longer be functionally equivalent. Therefore, it will be useful to characterize the expression characteristics and localization of the putative *UROD* isozymes.

The product of the reaction catalyzed by *UROD*, coproporphyrinogen III, is oxidized by *CPX* (step 8), which is encoded by two different genes in *C. rein-*

hardtii. EST sequences are available for both *CPX* genes (Tables I and III). The full-length sequence of *CPX1* was previously reported, and its gene product was purified and shown to be localized in the plastid (Hill and Merchant, 1995; Quinn et al., 1999). The deduced *CPX* protein sequences are compared in Supplemental Figure 5. The genomes of Arabidopsis and *T. pseudonana* also contain two potential *CPX* genes, while the *C. merolae* genome has a single *CPX* gene. The two *CPX* isoforms from Arabidopsis are very similar at the nucleotide level, which probably reflects a recent gene duplication. However, *CPX2* from Arabidopsis does not appear to encode a functional product since the ORF contains a frame shift (Santana et al., 2002). A single copy of the *CPX* gene was detected in other vascular plant genomes.

As noted earlier, *CPX* from vascular plants and *C. reinhardtii* is most similar to the mitochondrial enzyme from animals and fungi. The two *CPX* genes from *C. reinhardtii* have no intron positions in common, and the deduced amino acid sequences of the isozymes differ significantly (Supplemental Fig. 5); the amino acid sequences of the two *CPX* proteins from *T. pseudonana* are also very different. The *CPX1* isozymes of the two algae cluster with the single *CPX* proteins from vascular plants and *C. merolae* and the *CPX* homolog from the prasinophyte *Ostreococcus tauri*. By contrast, *C. reinhardtii* and *T. pseudonana* *CPX2* isoforms group in a separate cluster, positioned between the cyanobacterial and animal/fungal *CPX* clusters (Fig. 6). The same branching pattern could be reproduced with high bootstrap support ($n = 100$) by a maximum-likelihood analysis of the data set (data not shown). As there is a close relationship between algal *CPX2* and mitochondrial *CPX*, it will be important to establish the subcellular location(s) of *CPX2* in *C. reinhardtii*. Cyanobacteria of the genus *Nostoc* also have two *CPX* genes. However, these isoforms cluster within the cyanobacterial branch of the tree, suggesting that they are the result of a recent, local gene duplication that is restricted to a subgroup of the cyanobacteria.

Interestingly, vascular plants generally seem to contain two protoporphyrinogen IX oxidase (PPX) isozymes (step 9), which catalyze the step in Chl and heme biosynthesis immediately following *CPX*. In tobacco, one of these isozymes has been shown to be plastid specific, while the other was localized to mitochondria (Lermontova et al., 1997). The three algal genomes that we examined each contain a single *PPX* gene, encoding a protein that is most similar to the plastid-specific *PPX* from tobacco and the *PPX1* (At4g01690) from Arabidopsis (Table I).

Mg-chelatase (step 10) is situated at another important branch point in the tetrapyrrole biosynthetic pathway, catalyzing the committed step leading to Chl formation. This reaction has been recognized as an important target for regulation and has been the focus of several studies (e.g. see Walker and Willows, 1997; Beale, 1999). The active Mg-chelatase complex is a

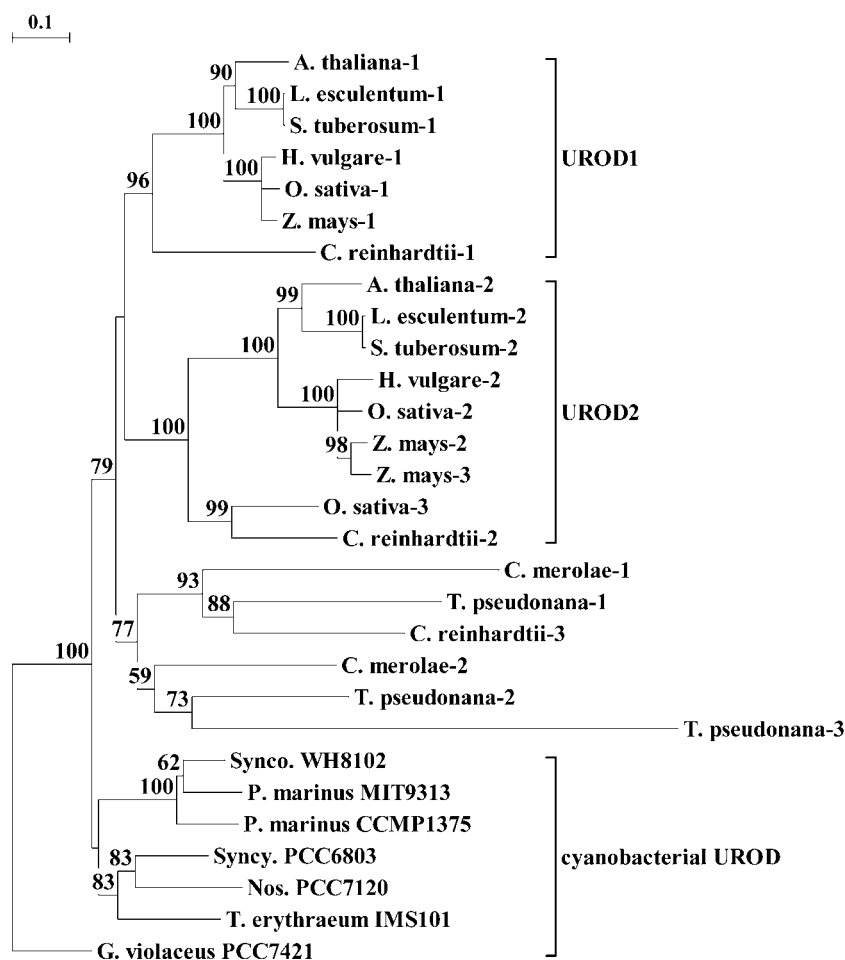


Figure 5. Neighbor-joining tree for UROD constructed from protein alignment (314 amino acid positions) of isozymes from *C. reinhardtii* with homologs from other algae, vascular plants, and cyanobacteria. The protein sequence from the cyanobacterium *Gloeobacter violaceus* PCC 7421 was used to root the tree. Bootstrap support higher than 50% is indicated at respective nodes ($n = 1,000$). Species and sequence accessions are as follows: *Arabidopsis* 1, NP_181581, and 2, NP_850587; *C. merolae* 1, CMP083C, and 2, CME194C; *G. violaceus* PCC 7421, NP_926823; *H. vulgare* 1, CAA58039, and 2, assembled from CB883221, AL508038, BG299928, BJ483280, BJ455669, AV834454; *L. esculentum* 1, assembled from BE462354, AI778274, AW624527, BI924627, BF112811, BI421978, and 2, assembled from BI928533, BI926422, BG124909, BI929528, BG129172, BM411365; *Nostoc* PCC 7120, NP_487949; *O. sativa* 1, AK070859, 2, AK106203, and 3, AK110601; *Prochlorococcus marinus* CCMP 1375, NP_875471; *P. marinus* MIT 9313, NP_894278; *S. tuberosum* 1, assembled from CV474374, BE921296, BQ047515, BE923279, BQ113938, BG097583, BM404774, and 2, assembled from CK276507, AW906474, BG097407, BE341237, CK276508, CK268172, BQ118921; *Synechococcus* WH 8102, NP_897588; *Synechocystis* PCC 6803, NP_442753; *T. pseudonana* 1, newV2.0.genewise.20.4.1 [thaps1:129030], 2, genewise.8.624.1 [thaps1:45777], and 3, assembled from unplaced WGS reads PQI71568.x1, SXZ37045.y1, PQI86858.y1, SXZ116998.y1, SXZ24083.x1, SXZ24793.x1, SXZ7875.x1, SXZ62420.y1, SXZ28894.y1, PQI126422.y1; *Trichodesmium erythraeum* IMS101, ZP_00326423; and *Z. mays* 1, O81220, 2, assembled from CD434648, CD435388, CD436605, and 3, assembled from CD442487, CD444385, CD977856, CA400912.

multimer composed of three different subunits termed CHLD, CHLH, and CHLI in eukaryotes and cyanobacteria, and bchD, bchH, and bchI in photosynthetic bacteria (they are required for bacteriochlorophyll synthesis). All subunits of the complex are highly conserved among photosynthetic organisms, with CHLI showing the highest sequence identity (Beale, 1999). In addition, the N-terminal half of CHLD exhibits significant similarity to the smaller CHLI subunit.

The *C. reinhardtii* genome contains two copies each of the *CHLI* and *CHLH* genes (Table I). Both *CHLI*

isogenes are expressed since there are several EST sequences in the *C. reinhardtii* database for each (Tables I and III). In addition, there is a full-length cDNA sequence for *CHLI1* (Lake and Willows, 2003). The core domains of *CHLI1* and *CHLI2* from *C. reinhardtii* show 62% identity and 82% similarity. *CHLI2* has a C-terminal extension of approximately 40 amino acids that is absent from the *CHLI* enzymes from other organisms; it may modify the function of this chelatase subunit (see Supplemental Fig. 6 and discussion below). Using the *CHLI1* and *CHLI2* sequences

Table III. Comparison of EST data and ENC values for ORFs in the *C. reinhardtii* genome encoding putative proteins involved in biosynthesis of chlorophylls (1–16) or carotenoids (21–43)

C. reinhardtii EST clones deposited in GenBank (<http://www.ncbi.nlm.nih.gov>) were grouped into three categories: (1) unstressed (clones of projects 874, 894, and 1,024 from Shrager et al., 2003; clones whose accession begins with "AV" from Asamizu et al., 1999, 2000); (2) stressed (projects 963, 1,031, and 3,510 [=1,115] from Shrager et al., 2003; projects 832 and 833); (3) other (projects 1,030, 3,511 (=1,112), and 925 from Shrager et al., 2003; clones whose accession begins with "BP" from Asamizu et al., 2004). EST clones with more than one sequence entry in GenBank (5'- and 3'-reads) were counted only once. The detailed frequencies of EST clones in each project are available as Supplemental Tables I and II. The ENC used by each ORF was calculated according to Wright (1990). For UROD, CPX, CHLI, and CHLH, the respective isogene that is represented by the majority of EST clones in cDNA libraries generated from unstressed cells and at the same time has the lowest ENC value is in bold.

Step	Gene Product	EST Clones				ENC
		Unstressed	Stressed	Other	Total	
1	GTS	0	0	0	0	29.4
2	GTR	5	13	5	23	32.4
3	GSA	6	3	5	14	25.7
4	ALAD	24	7	13	44	28.6
5	PBGD	11	20	4	35	28.0
6	UROS	3	4	2	9	30.6
7	UROD1	19	6	10	35	23.7
	UROD2	3	7	6	16	30.3
	UROD3	9	8	2	19	34.3
8	CPX1	14	4	8	26	26.1
	CPX2	3	2	0	5	35.2
9	PPX	1	0	0	1	30.5
10a	CHLD	6	2	3	11	30.9
10b	CHLI1	15	17	9	41	25.8
	CHLI2	6	0	3	9	31.9
10c	CHLH1	21	4	19	44	25.9
	CHLH2	0	0	0	0	33.8
11	PPMT	7	2	0	9	31.2
12	CHL27A	18	71	17	106	28.8
	CHL27B	9	2	1	12	24.8
14	LPOR	7	16	7	30	23.7
15	CAO	13	5	10	28	31.4
16	CHS	8	3	3	14	28.7
21	DXS	22	12	12	46	31.3
22	DXR	1	3	1	5	27.5
23	CMS	3	1	1	5	38.5
24	CMK	1	3	0	4	35.2
25	MCS	6	1	0	7	34.5
26	HDS	12	0	3	15	25.9
27	IDS	16	4	7	27	28.3
28	IDI	1	0	0	1	33.3
29	GGPS	2	8	3	13	32.6
30	PSY	8	1	9	18	40.0
31	PDS	5	6	6	17	29.7
32	ZDS	1	4	3	8	28.8
33	CRTISO	0	0	0	0	30.4
34	LCYB	7	3	6	16	32.7
35	LCYE	6	1	3	10	33.5
36	CHYB	0	0	0	0	31.4
37	CHYE	0	0	0	0	?
38	ZEP	5	2	3	10	30.3
42	BKT	1	0	0	1	39.5
43	GGR	37	5	0	57	26.5

from *C. reinhardtii*, we were able to assemble the respective genes from the WGS reads for *V. carteri*, and included the deduced protein sequences in our analyses.

In four other green algae, *Chlorella vulgaris* (Trebouxioiophyceae), *Mesostigma viride*, *Nephroselmis olivacea*

(both Prasinophyceae), and *Chaetosphaeridium globosum* (Charophyceae), unique genes encoding CHLI are located on the chloroplast genome; the plastome of *C. reinhardtii* does not harbor a CHLI gene. Moreover, the CHLI gene of four red algae, a euglenophyte, a

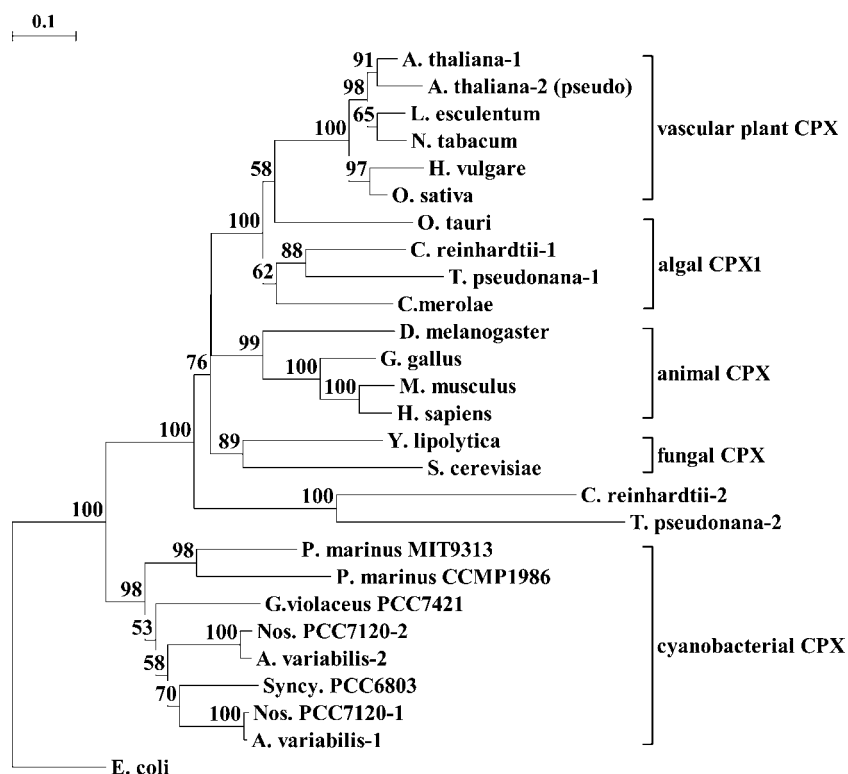


Figure 6. Neighbor-joining tree for CPX constructed from protein alignment (275 amino acid positions) of isozymes from *C. reinhardtii* with homologs from other algae, vascular plants, fungi, animals, and cyanobacteria. The protein sequence from the enterobacterium *E. coli* was used to root the tree. Bootstrap support higher than 50% is indicated at respective nodes ($n = 1,000$). Species and sequence accessions are as follows: *Anabaena variabilis* ATCC 29413 1, ZP_00159111, and 2, ZP_00158651; *Arabidopsis* 1, CAD12661, and 2 (pseudogene), NP_567256; *C. merolae*, CMO136C; *Drosophila melanogaster*, NP_524777; *Escherichia coli*, BAB36730; *Gallus gallus*, XP_416596; *Gloeobacter violaceus* PCC 7421, NP_926822; *Homo sapiens*, NP_000088; *H. vulgare*, CAA58037; *L. esculentum*, BT014254; *Mus musculus*, BAA03840; *N. tabacum*, CAA58038; *Nostoc* PCC 7120 1, NP_484694, and 2, NP_485400; *O. tauri*, AAS88901; *O. sativa*, XP_473852; *Prochlorococcus marinus* CCMP 1986, NP_893699; *P. marinus* MIT 9313, NP_895533; *S. cerevisiae*, AAA34529; *Synechocystis* PCC 6803, NP_440183; *T. pseudonana* 1, newV2.0.gene-wise.69.87.1 [thaps1:154535], and 2, newV2.0.grail.67.2.1 [thaps1:109130]; and *Yarrowia lipolytica*, XP_504574.

cryptophyte, a diatom, and a raphidophyte, is also encoded by the plastid genome.

Among the available sequences from vascular plants, we only identified two CHLI isoforms for *Arabidopsis*. These isozymes are 88% identical and 97% similar in the core region of the protein, suggesting that the two genes are the consequence of a recent duplication.

In phylogenetic analyses of CHLI from algae, vascular plants and cyanobacteria applying neighbor-joining (Fig. 7A) or maximum-likelihood methods (data not shown), CHLI1 of *C. reinhardtii* grouped with the CHLI proteins from other Chl *b*-containing organisms, i.e. the other green algae, *Euglena gracilis*, and vascular plants. Interestingly, CHLI2 from *C. reinhardtii* and *V. carteri* were well separated from plant and cyanobacterial CHLI proteins, making them somewhat unusual. As mentioned above, major regions of the CHLI proteins are highly conserved; therefore, only a limited number of informative amino acid positions are available for generating a phylogeny. We neglected to correct our analyses for substitution-rate heterogeneity, which explains the exceptionally short branch lengths of the neighbor-joining tree presented in Figure 7A. However, while the tree does not reflect true evolutionary distances, the branching pattern should not be affected.

Since *C. reinhardtii* and *V. carteri* were the only organisms for which two putative isoforms of CHLI were identified and the only algae in which the CHLI genes were located on the nuclear genome, the isozyms may have originated from a recent gene dupli-

cation, possibly at the base of the order Volvocales. This is supported by the observation that the two CHLI genes from *C. reinhardtii* share an intron position but have no intron sites in common with vascular plants CHLI genes. In opposition to this hypothesis, the remote position of CHLI2 in the phylogenetic groupings suggests that it may not be the result of a recent gene duplication. However, a gene duplication following the transfer of plastome-encoded CHLI to the nucleus may have relaxed the selective pressure on the isogenes (as a consequence of more than one gene copy), allowing for rapid divergence. As a consequence, the function of CHLI2 may be significantly different from that of CHLI1. This possibility is congruent with the finding that a number of highly conserved amino acids in all CHLI enzymes from a variety of distantly related photosynthetic organisms (i.e. all other organisms in the tree; Fig. 7B) are not conserved in the CHLI2 protein.

In summary, the CHLI isozymes from *C. reinhardtii* and *V. carteri* appear exceptional in two ways: (1) They are the only algal CHLI proteins known so far that are nucleus encoded, and (2) the second CHLI isoform present in these algae, CHLI2, has diverged to the extent that it may have significant differences in its activities relative to the highly conserved CHLI1.

In the case of the two potential CHLH proteins encoded on the *C. reinhardtii* genome, the sequence of CHLH1 is supported by considerable EST data, and a full-length sequence of the CHLH1 gene has been reported (Chekounova et al., 2001). The CHLH1 gene

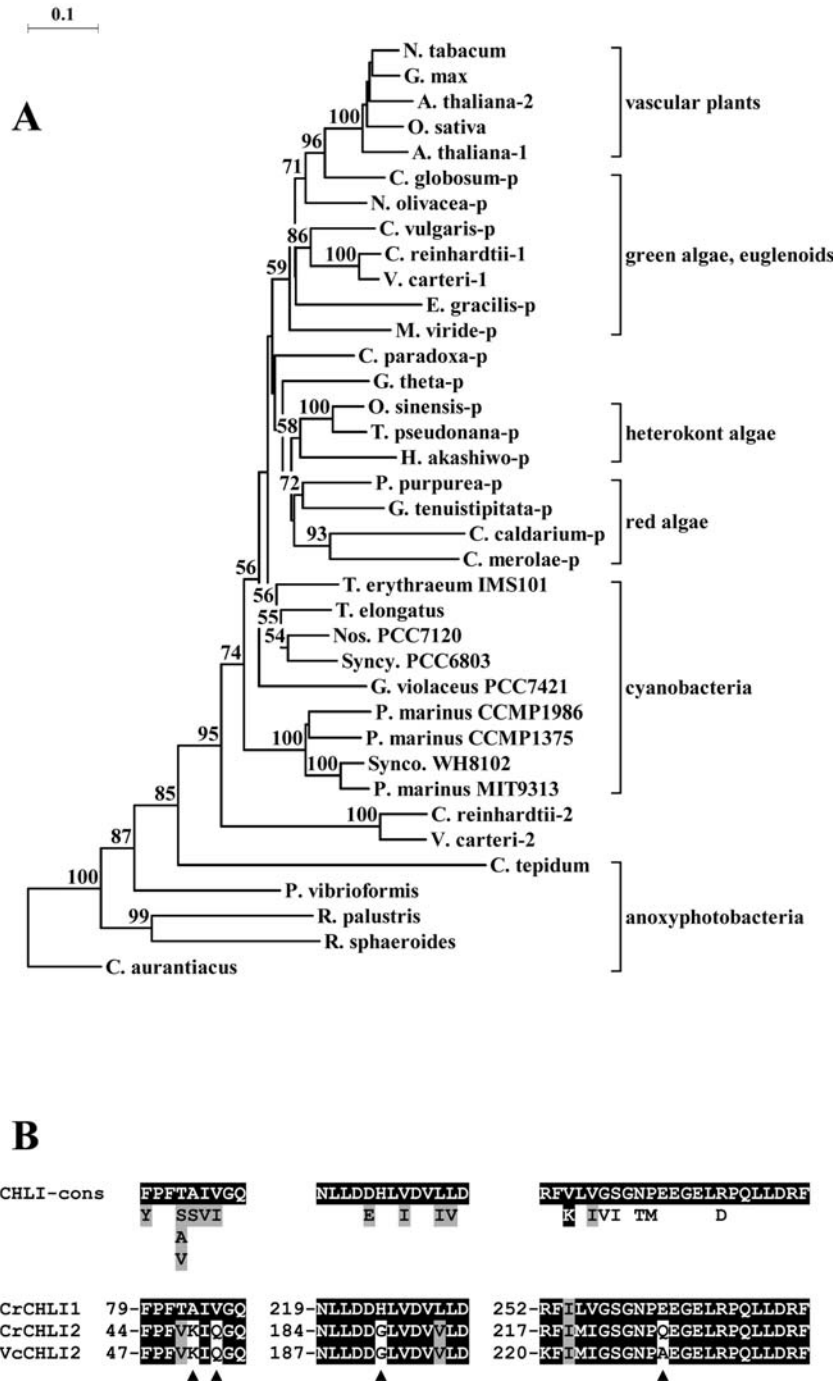


Figure 7. A, Neighbor-joining tree for CHLI constructed from protein alignment (314 amino acid positions) of isozyms from *C. reinhardtii* with homologs from vascular plants, diverse algae, cyanobacteria, and anoxygenic phototrophs. The suffix “p” in species names indicates that CHLI is encoded on the plastome of the respective species. The protein sequence from the bacterium *Chloroflexus aurantiacus* was used to root the tree. Bootstrap support higher than 50% is indicated at respective nodes ($n = 1,000$). B, The alignment of a consensus sequence deduced from all CHLI proteins in the tree with CHLI1 and CHLI2 isozyms from *C. reinhardtii* and *V. carteri*. Numbers at start of sequence blocks indicate position of the first shown amino acid in the full-length sequence. Black triangles label positions that are not conserved in the CHLI2 isozyms from the two Volvocales. Species and sequence accessions are as follows: Arabidopsis 1, NP_193583, and 2, NP_199405; *Chaetosphaeridium globosum*, AAM96508; *Chlorella vulgaris*, NP_045914; *Chlorobium tepidum*, AAM71667; *Chloroflexus aurantiacus*, AAG15216; *C. merolae*, BAC76114; *C. caldarium*, AAF12953; *Cyanophora paradoxa*, P48101; *E. gracilis*, CAA46470; *Gloeobacter violaceus* PCC 7421, BAC89655; *Glycine max*, BAA08291; *Gracilaria tenuistipitata*, AAT79784; *Guillardia theta*, AAC35636; *Heterosigma akashiwo*, CAA79971; *M. viride*, AAF43818; *N. olivacea*, NP_050837; *N. tabacum*, AAB97153; *Nostoc* PCC 7120, BAB77676; *O. sinensis*, CAA91677; *O. sativa*, AK070720; *P. purpurea*, AAC08280; *Prochlorococcus marinus* CCMP 1375,

encodes a predicted protein of 1,399 amino acids with over 65% identity to CHLH from Arabidopsis and *Synechocystis* PCC 6803 (Table I). A mutant of *C. reinhardtii* lacking CHLH1 is Chl deficient (Chekounova et al., 2001). For CHLH2, neither cDNA data nor a full-length genomic sequence is available. Database searches have revealed that the genomes of *C. merolae*, *T. pseudonana*, and several (but not all) cyanobacteria contain two distinct ORFs encoding potential CHLH isoforms that can be clustered into two groups (Fig. 8). By contrast, the vascular plants Arabidopsis and *O. sativa* contain a single CHLH gene. From alignments of the available CHLH protein sequences, we concluded that the *C. reinhardtii* CHLH2 gene model probably only includes the C-terminal half of the protein. However, we recovered genome fragments among the unplaced genomic reads from the *C. reinhardtii* genome database containing additional putative exons from CHLH2; the predicted polypeptide sequence encoded by these exons and the assembled C-terminal sequence of CHLH2 are aligned with CHLH1 in Supplemental Figure 7. (cyano)Bacteria have a protein designated CobN, which has significant sequence similarity to CHLH. This protein is a subunit of cobaltochelatase, an enzyme involved in cyanocobalamin synthesis. Cyanocobalamin does not appear to be synthesized by vascular plants and algae, and no enzymes in these organisms appear to rely on a cobalamin cofactor (Martens et al., 2002; Ravanel et al., 2004). The distribution pattern of CHLH2 among taxa suggests that the CHLH2 gene probably was lost during evolution of vascular plants; it might be the more ancient isoform of the Mg-chelatase H-subunit, somewhat more closely related to CobN than is CHLH1. This, however, needs to be substantiated by more detailed analyses with broader taxon sampling.

Although a function for CHLH2 has not been definitely established, it is likely to have a physiological function since it has been preserved in cyanobacteria, red algae, and diatoms. The protein may be expressed under specific conditions that require some modification of the Mg-chelatase activity/properties. If CHLH2 still functions in the association of Mg²⁺ with protoporphyrin IX, then known *chlH1* mutants of *C. reinhardtii* should be rescued by the introduction of CHLH2 expressed from a functional promoter. Such a system could be used to study function and regulation of CHLH2.

Finally, there are two isogenes coding for the Mg-protoporphyrin-IX monomethylester cyclase (CHL27; step 12) of *C. reinhardtii*, CHL27A and CHL27B, which have been studied extensively (Moseley et al., 2000, 2002; the latter reference contains an alignment of the two isoforms). The *Synechocystis* PCC 6803 genome also contains two ORFs (sll1214 and sll1874) with strong similarity to CHL27 from *C. reinhardtii* and vascular plants (Table I). *Nostoc* PCC7120 contains three genes encoding putative CHL27 homologs. The cyanobacterial isogenes are most similar to each other (Supplemental Fig. 8) and probably arose as the result of recent local gene duplications. A similar situation is observed with respect to the CHL27A and CHL27B of *C. reinhardtii*. For the vascular plants Arabidopsis, tomato, and *Z. mays*, we were only able to retrieve, from both genomic and EST databases, single sequences encoding CHL27. For *O. sativa*, there are EST sequences that suggest the occurrence of two different isoforms. However, while one of the isoforms groups with the vascular plant enzyme, the other is most closely related to the two proteins from *C. reinhardtii* and may represent contamination of the cDNA library. In red algae, CHL27 homologs are generally encoded on the plastid genome, as deduced from the completely sequenced plastomes of *C. merolae*, *Porphyra purpurea*, *Cyanidium caldarium*, and *Gracilaria tenuistipitata*. We could not detect sequences on the *T. pseudonana* genome encoding CHL27. Furthermore, there are no proteins with similarity to CHL27 encoded on the plastome of the diatom *Odontella sinensis*. These findings raise the possibility that the enzyme(s) catalyzing the formation of the cyclopentanone ring of the Chl molecule in diatoms is unrelated to CHL27.

Differential Expression of Genes Involved in Chl and Carotenoid Biosynthesis as Evaluated from EST Data and the ENC Used by the Genes

We wanted to learn more about expression levels of genes involved in the biosynthesis of Chl and carotenoids in *C. reinhardtii*. As a first step/approximation toward this goal, we analyzed the distribution of clones in the *C. reinhardtii* EST database for each enzyme of the pathway and calculated the ENC (Wright, 1990) in each of the ORFs. The results are compiled in Table III. Although most of the cDNA libraries from which sequences were generated by random sequencing had

Figure 7. (Continued.)

AAQ00191; *P. marinus* CCMP 1986, CAE19514; *P. marinus* MIT 9313, NP_894966; *Prosthecochloris vibrioformis*, CAB06299; *Rhodobacter sphaeroides*, AAF24287; *Rhodospseudomonas palustris*, CAE26948; *Synechococcus* WH 8102, NP_896809; *Synechocystis* PCC 6803, BAA17166; *T. pseudonana*, assembled from WGS reads SXZ77548.y1, SXZ216701.x1, SXZ196588.x1, SXZ132086.y1, SXZ11129.x2; *Thermosynechococcus elongatus* BP-1, BAC09063; *Trichodesmium erythraeum* IMS101, ZP_00326592; *V. carteri* CHLH1, assembled from WGS reads ABSY191795.g1, ABSY203211.g1, ABSY43528.b1, AOBN24164.y1, AOBN116614.y1, AOBN210642.y1, ABSY225494.y1, ABSY57889.g1, ABSY230014.y1, AOBN24164.x1, ABSY219172.g3, ABSY48314.g2; and CHLH2, assembled from ABSY137482.b2, ABSY28992.g1, ABSY75825.x3, ABSY84210.y3, ABSY4785.x1, ABSY134383.b1, ABSY42955.g1, AOBN146395.x1, ABSY121712.b1, ABSY47565.x1, AOBN218780.x1, AOBN193678.y1, ABSY207547.g1, ABSY79058.y1, ABSY100869.g1, AOBO4767.g2, AOBO4582.g2, AOBN193678.x3, AOBN21433.y1, ABSY89710.x3, ABSY188075.g1.

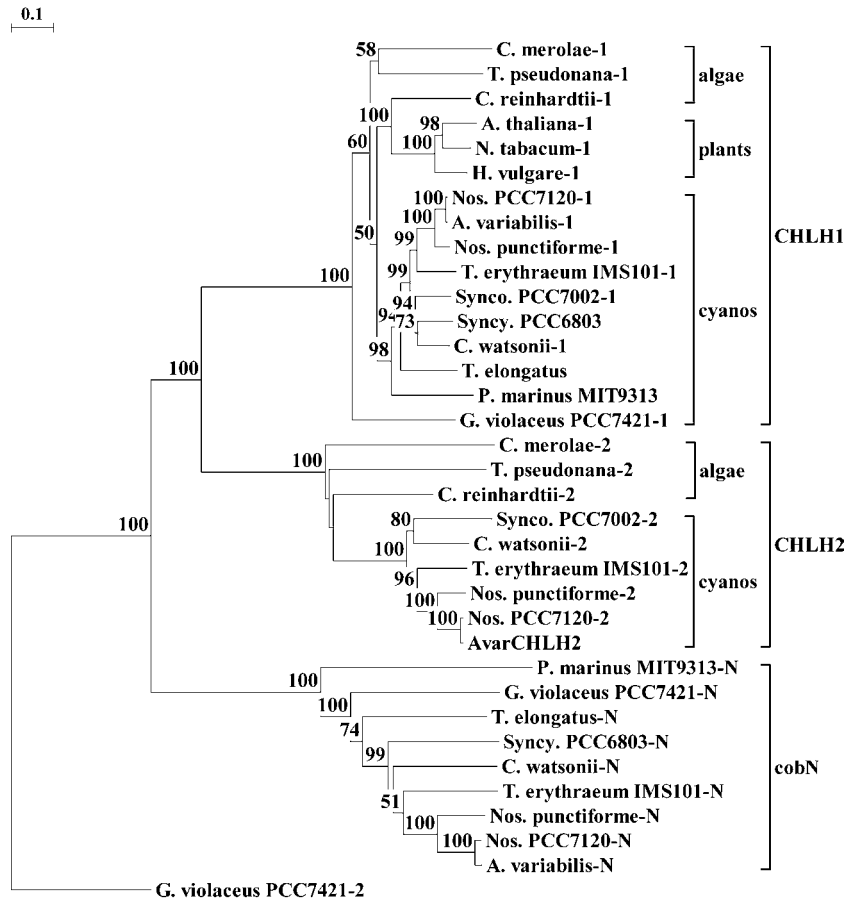


Figure 8. Neighbor-joining tree for CHLH and bacterial CobN constructed from protein alignment (1,093 amino acid positions) of isozyms from *C. reinhardtii* with homologs from vascular plants, algae, and cyanobacteria. The CHLH2 protein sequence from the cyanobacterium *Gloeobacter violaceus* was used to root the tree. Bootstrap support higher than 50% is indicated at respective nodes ($n = 1,000$). Species and sequence accessions are as follows: *Anabaena variabilis* ATCC 29413 CHLH1, ZP_00162059; CHLH2, ZP_00158885; CobN, ZP_00158224 (erroneous); *Arabidopsis*, CAA92802; *Crocospheera watsonii* WH 8501 CHLH1, ZP_00177805; CHLH2, ZP_00175982; CobN, ZP_00179369; *C. merolae* CHLH1, CMO212C; CHLH2, CMB093C; *Gloeobacter violaceus* PCC 7421 CHLH1, NP_925568; CHLH2, NP_926274; CobN, NP_926022; *H. vulgare*, AAK72401; *N. tabacum*, T01789; *Nostoc* PCC 7120 CHLH1, NP_488405; CHLH2, NP_488773; CobN, NP_485729; *Nostoc punctiforme* PCC 73102 CHLH1, ZP_00107632; CHLH2, ZP_00105988; CobN, ZP_00108621; *Prochlorococcus marinus* MIT 9313, NP_894647; CobN, NP_894559; *Synechococcus* PCC 7002 CHLH1, jmarq_32049|Contig051302-254; CHLH2, jmarq_32049|Contig051302-272; *Synechocystis* PCC 6803, NP_440360; CobN, NP_441516 + NP_440713; *T. pseudonana* CHLH1, newV2.0.genewise.28.59.1 [thaps1:168284]; CHLH2, grail.51.62.1 [thaps1:107764]; *Thermosynechococcus elongatus* BP-1, NP_681062; CobN, NP_681690; *Trichodesmium erythraeum* IMS101 CHLH1, ZP_00071825; CHLH2, ZP_00071317; and CobN, ZP_00325154.

been normalized, highly expressed genes still appear to be represented by greater clone frequency than genes with low expression levels; this is suggested by the large numbers of EST sequences for nucleus-encoded proteins involved in photosynthesis (e.g. the small subunits of ribulose-bisphosphate carboxylase, subunits of the two photosystems, apoproteins of the LHCs). Therefore, there is still some validity in using EST frequency and distributions as a first approximation of potential transcript levels under the conditions used to grow the cells for library construction.

The ENC used by an ORF to encode the 20 different amino acids from which a protein is synthesized can

theoretically vary between 20 (a single codon used for each amino acid) and 61 (all synonymous codons used for each amino acid; stop codons excluded; Wright, 1990). The evolution of codon bias within ORFs may be most consistently linked to the rate of translation of specific transcripts and the need to optimize protein production under specific environmental/developmental conditions. An ORF for an abundant protein may comprise a highly biased string of codons that pair with the most abundantly represented set of tRNAs in the cell (Ikemura, 1985). As a consequence, a low ENC value is mainly selected for at the level of translation. We are aware that other factors might

contribute to codon bias to some extent, such as the effect of specific codons on RNA secondary structure and genomic positional effects that might reflect the occurrence of chromosomal isochores. However, a detailed assessment of such effects will only be feasible after assembly of all of the linkage groups comprising the *C. reinhardtii* genome.

Although again an approximation, the ENC value may have the potential to provide hints about relative expression levels of isozymes. A correlation between the expression level of a gene and its codon usage bias has been reported for bacteria such as *E. coli* and *Salmonella typhimurium* (Bennetzen and Hall, 1982; Sharp and Li, 1987; dos Reis et al., 2003), the yeast *Saccharomyces cerevisiae* (Bennetzen and Hall, 1982; Sharp et al., 1986), the vascular plant *Arabidopsis* (Chiapello et al., 1998), and recently for *C. reinhardtii* (Naya et al., 2001).

The *C. reinhardtii* genome has a high GC bias and an average ENC value, based on 663 *C. reinhardtii* ORFs registered in the codon usage database (<http://www.kazusa.or.jp/codon/>), of 32.5. In some highly expressed genes, however, the ENC value is close to its lower limit of 20. For example, the *RBCS1* gene, which encodes a small-subunit isozyme of ribulose-bisphosphate carboxylase, and the gene encoding glyceraldehyde-3-phosphate dehydrogenase have ENC values of 22.7 and 24.1, respectively.

Genes encoding putative isozymes of the Chl biosynthetic pathway of *C. reinhardtii* exhibit significant differences with respect to both EST frequencies and ENC values (Table III). In the cases of the *UROD*, *CPX*, *CHL1*, and *CHLH* isogenes of *C. reinhardtii*, those that are represented by the majority of EST clones in cDNA libraries generated from cells grown under favorable conditions (unstressed) also have especially low ENC values. The protein products of such isogenes (*UROD1*, *CPX1*, *CHL11*, *CHLH1*; highlighted by bold letters in Table III) are probably more abundant than those of the isogenes with lower EST representation and higher ENC values, making it likely they represent the major isozymes involved in Chl and heme biosynthesis. For *CPX1* (Quinn et al., 1999) and *CHLH1* (Chekounova et al., 2001), this is congruent with previously reported data.

Other genes involved in Chl biosynthesis with remarkably low ENC values are *GSA*, *CHL27B*, and *LPOR*, while among carotenogenic genes only *HDS* has an ENC value of below 26. To further substantiate that the low ENC values for these genes are related to translational selection, we analyzed the codon bias for each of these genes. The most highly selective codon usage was for Pro (CCC), Thr (ACC), Gly (GGC), Leu (CTG), and Arg (CGC). This finding agrees with the analyses of other highly expressed genes in *C. reinhardtii*, as reported by Naya et al. (2001).

On average, the ENC values of genes of the Chl biosynthetic pathway are considerably lower than those of carotenogenic genes (median of 28.0 compared to 32.4 when considering only the lowest ENC in the case of isogenes, and counting *GGR* as a Chl

biosynthetic gene; $P < 0.01$; Mann-Whitney U test). This finding might reflect a more urgent need to regulate expression of genes involved in Chl biosynthesis since the intermediates of this pathway can be extremely toxic.

Light Is Involved in Controlling Genes Encoding Proteins Required for Chl and Carotenoid Synthesis

Light is one critical environmental factor that controls the synthesis of both Chl and carotenoids. In vascular plants, ALA synthesis is regulated by a range of signals, including photoreceptors (McCormac et al., 2001), circadian rhythms (Kruse et al., 1997), and the developmental status of the plastid (Kumar et al., 1999). The expression of *GSA*, which converts Glu-1-semialdehyde to ALA, has been shown to be regulated by light in vascular plants and *C. reinhardtii* (Ilag et al., 1994; Im et al., 1996; Im and Beale, 2000). While *C. reinhardtii* synthesizes Chl in the dark, Chl synthesis increases in the light (Janero and Barnett, 1982), and light inputs (Johanningmeier and Howell, 1984) may help tailor the structure and activity of the photosynthetic apparatus for photoautotrophic growth. The carotenoid biosynthetic pathway is also regulated by several environmental factors (Bouvier et al., 1996; Audran et al., 1998; Steinbrenner and Linden, 2001).

Some of the specific enzymes involved in Chl biosynthesis are encoded by either two or three distinct genes. We used quantitative PCR (qPCR) to compare expression of the different members of these multigene families in cells maintained in the dark, with their expression following exposure of the cells to light (Fig. 9A). Furthermore, we included other potential regulatory targets from both the Chl and the carotenoid biosynthetic pathway in our analysis. Very low fluence white light (VLFL; $0.01 \mu\text{mol photon m}^{-2} \text{s}^{-1}$) was used for these experiments to eliminate the influence of changes in the redox state of the cell as a consequence of photosynthetic electron transport. While the transcripts for a number of the genes increased following exposure to VLFL, transcript level were generally higher after 2 h than after 4 h of VLFL treatment.

As shown in Figure 9A, the transcripts for both *GSA* and *ALAD* markedly increase (approximately 15-fold for *GSA* and 7-fold for *ALAD*) after a 2 h exposure to VLFL, supporting the previous findings that indicated that these genes were under light control (Matters and Beale, 1994). Since the light intensity used in these experiments was too low to affect photosynthetic activity, it is likely that a specific photoreceptor(s) is involved in the light-dependent increase in transcript levels. There was a 4- to 6-fold increase in *UROD1* transcript levels after exposure of *C. reinhardtii* to VLFL for 2 h. However, transcripts encoding *UROD2* and *UROD3* showed no significant increase following VLFL exposure. This observation further supports the preliminary conclusion drawn from the EST and ENC analyses that *UROD1* is the major isozyme involved

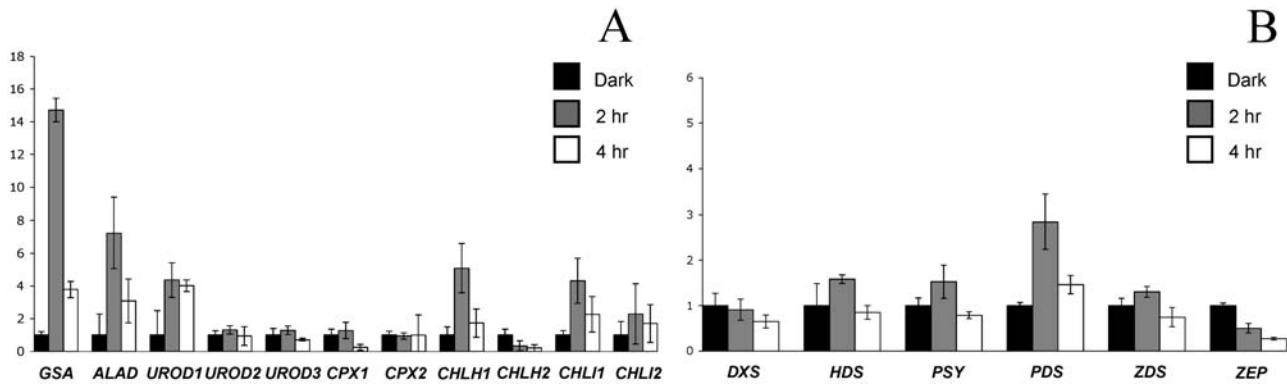


Figure 9. Bar graphs showing changes in transcript levels for genes involved in Chl (A) and carotenoid (B) biosynthesis in *C. reinhardtii* strain CC124 following exposure to VLFL, as analyzed by qPCR. The responses, shown as the change (n -fold) relative to RNA from dark grown cells, at the different times following exposure to VLFL are given as different bar fill patterns, as indicated on the graph. The graph shows relative expression levels of each gene normalized to the *CBLP* gene (the level of this transcript remains constant over the course of the experiment). The results show the mean and SD for triplicate qPCR results.

in the biosynthesis of Chl and heme. The amount of UROD mRNA and protein were previously shown to increase following illumination of barley seedlings (Mock et al., 1995).

CPX is encoded by *CPX1* and *CPX2* isogenes in *C. reinhardtii*. Expression of *CPX1* has been shown to be influenced by both copper (Cu) and oxygen levels (Quinn et al., 1999; Moseley et al., 2000). There are three different-sized mRNAs transcribed from the *CPX1* gene; the shortest becomes very abundant in Cu-deprived cells, while levels of the other forms of the transcript are insensitive to Cu availability (Quinn et al., 1999). EST/ENC analyses suggest that *CPX1* encodes the more abundant isoform in *C. reinhardtii*. However, based on qPCR, the expression levels of both *CPX1* and *CPX2* were very low, with little change following exposure of dark-grown cells to VLFL. In barley, CPX mRNA levels did not significantly differ in etiolated and greening barley leaves, but did change as the seedling developed (Kruse et al., 1995). Therefore, while the absorption of light by specific photoreceptors may not play a big role in controlling CPX expression, these genes appear to be controlled by other signals, including Cu availability, local oxygen levels, and perhaps also the redox state of the cell.

In the case of the subunits of the Mg-chelatase complex, the *CHLH1* transcript of *C. reinhardtii* increases in VLFL (approximately 5-fold), while the level of the *CHLH2* transcript is very low in both the dark and VLFL (no significant difference in the transcript level in dark and VLFL). Mutant lines of *O. sativa* in which the *CHLH* gene was disrupted are chlorotic (Jung et al., 2003), as are tobacco antisense lines (Papenbrock et al., 2000). It was previously shown that the level of the *CHLH1* transcript of *C. reinhardtii* is influenced by light and that a *chlH1* mutant in this alga leads to a chlorotic, high light-sensitive phenotype (Chekounova et al., 2001). Interestingly, *CHLH* antisense transgenic tobacco plants do not accumulate protoporphyrin IX and exhibit reduced levels of mRNA encoding a number of enzymes involved in the early steps of

Chl biosynthesis (Papenbrock et al., 2000). These results suggest a mechanism in which metabolic intermediates in Chl biosynthesis can modulate expression of genes encoding enzymes that function in earlier steps in the pathway. By contrast, a *chlH1* mutant of *C. reinhardtii* does not exhibit decreased levels of mRNAs encoding at least some enzymes that function early in Chl biosynthesis, although it does have a chlorotic phenotype. The level of ALAD mRNA remained high, and notably the levels of the transcripts for the CHLI and D subunits were comparable to those of wild-type cells (Chekounova et al., 2001). Hence, there are aspects of control of Chl biosynthesis in *C. reinhardtii* that appear to be different from those of vascular plants. However, in spite of the finding that a *chlH1* mutant of *C. reinhardtii* was chlorotic (Chekounova et al., 2001), the expression of *CHLH2* may be required for tuning the level of Chl synthesis under specific environmental conditions. Furthermore, transcripts encoded by both *CHLI* genes of *C. reinhardtii* were elevated following exposure to VLFL, although the increase in *CHLI1* mRNA was higher than that of *CHLI2*. Light induction of *CHLI* has been previously observed in soybean (Nakayama et al., 1995).

Carotenoid biosynthesis has also been shown to be regulated by light in both *Arabidopsis* and *C. reinhardtii* (Bohne and Linden, 2002; Woitsch and Romer, 2003; Botella-Pavia et al., 2004; Liu et al., 2004). However, in many cases it is still not clear whether the light requirement reflects contributions to regulation from photoreceptors, redox conditions of the cell, or both. As shown in Figure 9B, we used qPCR to examine expression of many genes involved in the synthesis of carotenoids. The only transcript that clearly increases (of the genes tested) upon exposure of *C. reinhardtii* to VLFL is that of *PDS*, suggesting that control of this gene is at least in part a consequence of the action of a specific photoreceptor. Increases in *HDS*, *PSY*, and *ZDS* transcript levels were relatively small, although they do appear to be significant. In *Arabidopsis*, *DXS*, *HDR* (designated *IDS* in Table II

and Fig. 2), and *PSY* transcripts increased during deetiolation of seedlings (Botella-Pavia et al., 2004). However, these experiments did not use the extremely low light intensities that were employed in the experiments presented in this article.

In *C. reinhardtii*, blue light elicits an increase in the level of *GSA* mRNA in cultures grown under conditions of light:dark synchronization (Matters and Beale, 1995; Herman et al., 1999). The signal transduction events involved in photoregulation of the *GSA* gene are proposed to include activation of a heterotrimeric G-protein and phospholipase C, an increase in cytosolic Ca^{2+} levels, and activation of calmodulin and calmodulin-kinase (Im et al., 1996; Im and Beale, 2000). Light regulation of *GSA* also requires that the cells have access to adequate carbon (acetate) and nitrogen sources in the growth medium (Im et al., 1996), suggesting that signals from other metabolic pathways influence expression of *GSA* and interact either directly or indirectly with light signals. Induction of *GSA*, *ALAD*, *UROD1*, *CHLH1*, *CHL11*, and *CHL12* by VLFL suggests that these genes may have common regulatory mechanisms. One candidate for the photoreceptor involved in controlling genes encoding the enzymes responsible for Chl biosynthesis is PHOT1, a flavin-binding photoreceptor recently implicated in blue light induction of gametogenesis in *C. reinhardtii* (Huang et al., 2002). Preliminary studies using a PHOT1 RNAi strain suggest that VLFL-stimulated expression of *GSA* is under the control of PHOT1 (C.S. Im, C.F. Beck, and A.R. Grossman, unpublished data). While most transcripts encoding proteins involved in carotenoid biosynthesis (that we have tested) are not strongly regulated by VLFL, the *PDS* transcript shows a 3-fold increase after 2 h in VLFL, raising the possibility that *PDS* may also be under the control of PHOT1.

CONCLUDING REMARKS

The results discussed in this article demonstrate that genomic analyses of biosynthetic pathways in *C. reinhardtii* can reveal the occurrence of families of genes for a specific biosynthetic step in the pathway, phylogenetic relationships of the deduced protein sequences with those of other organisms, the sequences that target these proteins to the chloroplast, and the occurrence of specific conserved domains in plant and algal polypeptides not present in their cyanobacterial counterparts. The deduced amino acid sequences of proteins involved in Chl and carotenoid biosynthesis of *C. reinhardtii* also point to some intriguing differences among the algal, cyanobacteria, and vascular plant proteins with respect to both structure and regulation of activity (e.g. for the *CAO* gene). Furthermore, the genomic and cDNA information demonstrates that some of the enzymes involved in Chl biosynthesis are encoded by gene families in *C. reinhardtii*, and that differential regulation of specific members of these families may provide a mechanism by which the alga can acclimate to different light

conditions, and perhaps to other environmental conditions. Additional light studies using specific wavelengths of light as well as high-intensity light, coupled with microarray analyses, are beginning, and will likely provide us with information on the role of specific photoreceptors and redox levels in the control of pigment biosynthesis in *C. reinhardtii*.

MATERIALS AND METHODS

Strains and Culture Condition

The *Chlamydomonas reinhardtii* wild-type strain (parental strain) CC124 was used for all experiments presented in this study. The cells were grown in Tris acetate phosphate medium at a moderate/low light intensity ($45 \mu\text{mol photon m}^{-2} \text{s}^{-1}$) to a density of 5×10^5 cells mL^{-1} , and then transferred to the dark for 24 h before exposure to light. Light treatments were performed with white LED (RL5-W6030; 6000mcd; Super Bright LEDs, St. Louis) at very low intensity ($0.01 \mu\text{mol photon m}^{-2} \text{s}^{-1}$).

Gene Identification and Sequence Analysis

To identify specific genes in the *C. reinhardtii* genome, either the sequence information from GenBank for previously isolated and characterized genes, or the sequence information from other organisms (mainly *Arabidopsis* [*Arabidopsis thaliana*] and cyanobacteria) derived from cDNA or genomic DNA sequence information, was used to perform BLAST (Altschul et al., 1997) alignments against the *C. reinhardtii* genome sequence (<http://genome.jgi-psf.org/chlre2>) and EST database (<http://www.chlamy.org/search.html>). Candidate orthologous and paralogous predicted proteins were aligned with each other using ClustalW (Thompson et al., 1994). Alignments were manually improved and documented using BioEdit software (Hall, 1999).

Identification of putative targeting signals was performed using the prediction tools TargetP (Nielsen et al., 1997; Emanuelsson et al., 2000; <http://www.cbs.dtu.dk/services/TargetP/>), ChloroP (Emanuelsson et al., 1999; <http://www.cbs.dtu.dk/services/ChloroP/>), Predotar (Small et al., 2004; <http://genoplante.info.infobiogen.fr/predotar/>), and iPSORT (Bannai et al., 2002; <http://biocaml.org/ipsort/iPSORT/>). Secondary structure of CAO sequences was predicted by PSIPRED (McGuffin et al., 2000; <http://bioinf.cs.ucl.ac.uk/psipred/>), and putative transmembrane helices in BKT were predicted by DAS (Cserzo et al., 1997; <http://www.sbc.su.se/~miklos/DAS/>) and TMHMM (Krogh et al., 2001; <http://www.cbs.dtu.dk/services/TMHMM-2.0/>).

Genes coding for similar proteins from other organisms were retrieved by BLAST searches of GenBank genomes, nucleotide and EST entries, the current versions of the genomes from the red alga *Cyanidioschyzon merolae* (Matsuzaki et al., 2004; <http://merolae.biol.s.u-tokyo.ac.jp/>), the diatom *Thalassiosira pseudonana* (Armbrust et al., 2004; <http://genome.jgi-psf.org/thaps1/thaps1.home.html>), and the WGS database for the green alga *Volvox carteri* in the National Center for Biotechnology Information trace archive (<http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml>). Assembly of mRNA and gene sequences from EST and WGS reads was done in BioEdit.

Phylogenetic analyses were performed using the software packages Treecon (Van de Peer and De Wachter, 1994) and PHYLIP (Felsenstein, 1989). From the protein alignments, neighbor-joining trees based on Poisson-corrected distances were constructed in Treecon 1.3b, while maximum-likelihood analyses applying the JTT substitution matrix were performed with the PAML module of PHYLIP 3.61.

To determine EST frequencies of each gene, EST clones were identified by BLAST of *C. reinhardtii* EST entries in GenBank (as of August 2004) with putative full-length cDNA sequences of the respective genes. For calculation of ENC data, the codon usage frequencies of each ORF were analyzed with the program SPIN, part of the Staden package (Staden, 1996; <http://staden.sourceforge.net/>), and the results were imported into Microsoft EXCEL (Redmond, WA) and used to calculate ENC values according to Wright (1990) in a custom-made spreadsheet.

Sequencing of cDNA

For some genes, additional cDNA sequencing was performed. In the *C. reinhardtii* EST database, we identified full-length clones encoding putative

CAO (AV626430), HDS (AV626792), LCYB (AV641959), and BKT (1024014H04) proteins, which were made available to us by the Kazusa DNA Research Institute (Chiba, Japan; Asamizu et al., 1999, 2000) and the Stanford Genome Technology Center (Stanford, CA; Shragar et al., 2003), respectively. In addition, we amplified a full-length sequence encoding CHYB from a recombinant λ ZAPII cDNA library using the Expand Long Template PCR system (Roche Applied Science, Indianapolis) as described elsewhere (Im and Grossman, 2002). Sequencing of cDNA and PCR fragments was performed on both strands using Big Dye sequencing reagents (Amersham Pharmacia Biotech, Piscataway, NJ); the fluorescent fragments were resolved on an ABI310 capillary sequencer. GenBank accessions for the full sequences are as follows: CAO, AY860816; HDS, AY860817; LCYB, AY860818; BKT, AY860820; and CHYB, AY860819.

RNA Isolation and RNA-Blot Analysis

Total RNA was isolated from cells using TRIZOL reagent (38% phenol, 0.8 M guanidine thiocyanate, 0.4 M ammonium thiocyanate, 0.1 M sodium acetate, pH 5, and 5% glycerol) containing 0.2 volumes of chloroform. The cells were lysed by suspension in the TRIZOL reagent, and nucleic acid in the aqueous layer was precipitated by adding 0.5 volumes of isopropanol, 0.5 volumes of 0.8 M sodium citrate/1.2 M NaCl. The RNA precipitate was allowed to form at 4°C for 4 h before it was pelleted by centrifugation at 10,000g for 30 min, washed with 70% ethanol, air-dried, and dissolved in sterile distilled water.

qPCR

Isolated total RNA was treated with RNase-free DNase I (Ambion, Austin, TX), followed by phenol:chloroform extraction. For cDNA synthesis, 1 μ g of DNase I-treated total RNA was reverse transcribed and amplified using the Superscript II kit (Invitrogen, Carlsbad, CA), as described by the manufacturer. qPCR was performed using the DyNamo Hot Star SYBR Green qPCR kit (MJ Research, Waltham, MA) and analyzed by the Opticon 2 real-time system (MJ Research). Cycling conditions included an initial incubation at 95°C for 10 s, followed by 40 cycles of 94°C for 10 s, 55°C for 15 s, and 72°C for 10 s. Each of the PCR assays was performed in triplicate. The relative expression ratio of target gene was calculated based on the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001). *CBLP* gene was used as control gene, and each primer was designed by Primer3 software (Rozen and Skaletsky, 2000; http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) to distinguish the different isozymes. Primers were designed to have a melting temperature between 58°C and 60°C, and an optimal length of 20 nucleotides. The GC content was held between 20% and 80% with no 3' GC clamp. The target amplicon for each sequence was designed to be between 150 and 200 nucleotides with an optimal melting temperature of 85°C.

Sequence data from this article have been deposited with the EMBL/GenBank data libraries under accession numbers AY860816 to AY860820.

ACKNOWLEDGMENTS

We thank Dan Rokhsar and Diego Martinez at JGI and members of the *Chlamydomonas* Genome Consortium for helping to develop the tools and infrastructure for securing and examining *C. reinhardtii* cDNA and genomic information, and for providing stimulating discussions and valuable insights. The supply of EST clones by the Kazusa DNA Research Institute (Chiba, Japan) and the Stanford Genome Technology Center (Stanford, CA) is gratefully acknowledged. We also are grateful to two anonymous reviewers for critically reading the manuscript and for further suggestions.

Received November 5, 2004; returned for revision February 3, 2005; accepted February 8, 2005.

LITERATURE CITED

Al-Babili S, Huguency P, Schledz M, Welsch R, Frohnmeyer H, Laule O, Beyer P (2000) Identification of a novel gene coding for neoxanthin synthase from *Solanum tuberosum*. *FEBS Lett* **485**: 168–172

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402
- Anwaruzzaman M, Chin BL, Li X-P, Lohr M, Martinez DA, Niyogi KK (2004) Genomic analysis of mutants affecting xanthophyll biosynthesis and regulation of photosynthetic light harvesting in *Chlamydomonas reinhardtii*. *Photosynth Res* **82**: 265–276
- Arabidopsis Genome Initiative (2001) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815
- Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, et al (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**: 79–86
- Asamizu E, Miura K, Kucho K, Inoue Y, Fukuzawa H, Ohyama K, Nakamura Y, Tabata S (2000) Generation of expressed sequence tags from low-CO₂ and high-CO₂ adapted cells of *Chlamydomonas reinhardtii*. *DNA Res* **7**: 305–307
- Asamizu E, Nakamura Y, Sato S, Fukuzawa H, Tabata S (1999) A large scale structural analysis of cDNAs in a unicellular green alga, *Chlamydomonas reinhardtii*. I. Generation of 3433 non-redundant expressed sequence tags. *DNA Res* **6**: 369–373
- Audran C, Borel C, Frey A, Sotta B, Meyer C, Simonneau T, Marion-Poll A (1998) Expression studies of the zeaxanthin epoxidase gene in *Nicotiana glumbaginifolia*. *Plant Physiol* **118**: 1021–1028
- Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* **18**: 298–305
- Baroli I, Do AD, Yamane T, Niyogi KK (2003) Zeaxanthin accumulation in the absence of a functional xanthophyll cycle protects *Chlamydomonas reinhardtii* from photooxidative stress. *Plant Cell* **15**: 992–1008
- Baroli I, Niyogi KK (2000) Molecular genetics of xanthophyll-dependent photoprotection in green algae and plants. *Philos Trans R Soc Lond B Biol Sci* **355**: 1385–1394
- Beale SI (1999) Enzymes of chlorophyll biosynthesis. *Photosynth Res* **60**: 43–73
- Bennetzen JL, Hall BD (1982) Codon selection in yeast. *J Biol Chem* **257**: 3026–3031
- Bohne F, Linden H (2002) Regulation of carotenoid biosynthesis genes in response to light in *Chlamydomonas reinhardtii*. *Biochim Biophys Acta* **1579**: 26–34
- Botella-Pavia P, Besumbes O, Phillips MA, Carretero-Paulet L, Boronat A, Rodriguez-Concepcion M (2004) Regulation of carotenoid biosynthesis in plants: evidence for a key role of hydroxymethylbutenyl diphosphate reductase in controlling the supply of plastidial isoprenoid precursors. *Plant J* **40**: 188–199
- Bougri O, Grimm B (1996) Members of a low-copy number gene family encoding glutamyl-tRNA reductase are differentially expressed in barley. *Plant J* **9**: 867–878
- Bouvier F, d'Harlingue A, Backhaus RA, Kumagai MH, Camara B (2000) Identification of neoxanthin synthase as a carotenoid cyclase paralog. *Eur J Biochem* **267**: 6346–6352
- Bouvier F, d'Harlingue A, Huguency P, Marin E, Marion-Poll A, Camara B (1996) Xanthophyll biosynthesis. Cloning, expression, functional reconstitution, and regulation of beta-cyclohexenyl carotenoid epoxidase from pepper (*Capsicum annuum*). *J Biol Chem* **271**: 28861–28867
- Breitenbach J, Misawa N, Kajiwara S, Sandmann G (1996) Expression in *Escherichia coli* and properties of the carotene ketolase from *Haemato-coccus pluvisialis*. *FEMS Microbiol Lett* **140**: 241–246
- Bugos RC, Hieber AD, Yamamoto HY (1998) Xanthophyll cycle enzymes are members of the lipocalin family, the first identified from plants. *J Biol Chem* **273**: 15321–15324
- Chekounova E, Voronetskaya V, Papenbrock J, Grimm B, Beck CF (2001) Characterization of *Chlamydomonas* mutants defective in the H subunit of Mg-chelatase. *Mol Genet Genomics* **266**: 363–373
- Chiappello H, Lisacek F, Caboche M, Henaut A (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* **209**: GC1–GC38
- Cornah JE, Terry MJ, Smith AG (2003) Green or red: What stops the traffic in the tetrapyrrole pathway? *Trends Plant Sci* **8**: 224–230
- Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng* **10**: 673–676

- Cunningham FX Jr, Gantt E (1998) Genes and enzymes of carotenoid biosynthesis in plants. *Annu Rev Plant Physiol Plant Mol Biol* **49**: 557–583
- Cunningham FX Jr, Gantt E (2000) Identification of multi-gene families encoding isopentenyl diphosphate isomerase in plants by heterologous complementation in *Escherichia coli*. *Plant Cell Physiol* **41**: 119–123
- Cunningham FX Jr, Pogson B, Sun Z, McDonald KA, DellaPenna D, Gantt E (1996) Functional analysis of the beta and epsilon lycopene cyclase enzymes of *Arabidopsis* reveals a mechanism for control of cyclic carotenoid formation. *Plant Cell* **8**: 1613–1626
- Demmig-Adams B (1990) Carotenoids and photoprotection in plants: a role for the xanthophyll zeaxanthin. *Biochim Biophys Acta* **1020**: 1–24
- Demmig-Adams B, Gilmore AM, Adams WW III (1996) In vivo functions of carotenoids in higher plants. *FASEB J* **10**: 403–412
- dos Reis M, Wernisch L, Savva R (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res* **31**: 6976–6985
- Eggink LL, LoBrutto R, Brune DC, Brusslan J, Yamasato A, Tanaka A, Hooper JK (2004) Synthesis of chlorophyll *b*: localization of chlorophyllide *a* oxygenase and discovery of a stable radical in the catalytic subunit. *BMC Plant Biol* **4**: 5
- Elrad D, Grossman AR (2004) A genome's-eye view of the light-harvesting polypeptides of *Chlamydomonas reinhardtii*. *Curr Genet* **45**: 61–75
- Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300**: 1005–1016
- Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci* **8**: 978–984
- Espineda CE, Linford AS, Devine D, Brusslan JA (1999) The AtCAO gene, encoding chlorophyll *a* oxygenase, is required for chlorophyll *b* synthesis in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **96**: 10507–10511
- Felsenstein J (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5**: 164–166
- Formaggio E, Cinque G, Bassi R (2001) Functional architecture of the major light-harvesting complex from higher plants. *J Mol Biol* **314**: 1157–1166
- Frank HA, Brudvig GW (2004) Redox functions of carotenoids in photosynthesis. *Biochemistry* **43**: 8607–8615
- Frank HA, Cogdell RJ (1993) The photochemistry and function of carotenoids in photosynthesis. In A Young, G Britton, eds, *Carotenoids in Photosynthesis*. Chapman and Hall, London, pp 253–326
- Franzen LG, Rochaix JD, von Heijne G (1990) Chloroplast transit peptides from the green alga *Chlamydomonas reinhardtii* share features with both mitochondrial and higher plant chloroplast presequences. *FEBS Lett* **260**: 165–168
- Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92–100
- Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. *Science* **283**: 1476–1481
- Grossman A (2000) *Chlamydomonas reinhardtii* and photosynthesis: genetics to genomics. *Curr Opin Plant Biol* **3**: 132–137
- Grossman AR, Harris EE, Hauser C, Lefebvre PA, Martinez D, Rokhsar D, Shrager J, Silflow CD, Stern D, Vallon O, et al (2003) *Chlamydomonas reinhardtii* at the crossroads of genomics. *Eukaryot Cell* **2**: 1137–1150
- Grossman AR, Lohr M, Im CS (2004) *Chlamydomonas reinhardtii* in the landscape of pigments. *Annu Rev Genet* **38**: 119–173
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**: 95–98
- Harris EH (2001) *Chlamydomonas* as a model organism. *Annu Rev Plant Physiol Plant Mol Biol* **52**: 363–406
- Havaux M, Niyogi KK (1999) The violaxanthin cycle protects plants from photooxidative damage by more than one mechanism. *Proc Natl Acad Sci USA* **96**: 8762–8767
- Herman CA, Im CS, Beale SI (1999) Light-regulated expression of the *gsa* gene encoding the chlorophyll biosynthetic enzyme glutamate 1-semialdehyde aminotransferase in carotenoid-deficient *Chlamydomonas reinhardtii* cells. *Plant Mol Biol* **39**: 289–297
- Hill KL, Merchant S (1995) Coordinate expression of coproporphyrinogen oxidase and cytochrome *c_c* in the green alga *Chlamydomonas reinhardtii* in response to changes in copper availability. *EMBO J* **14**: 857–865
- Hirschberg J (2001) Carotenoid biosynthesis in flowering plants. *Curr Opin Plant Biol* **4**: 210–218
- Huang K, Merkle T, Beck CF (2002) Isolation and characterization of a *Chlamydomonas* gene that encodes a putative blue-light photoreceptor of the phototropin family. *Physiol Plant* **115**: 613–622
- Huguene P, Badillo A, Chen HC, Klein A, Hirschberg J, Camara B, Kuntz M (1995) Metabolism of cyclic carotenoids: a model for the alteration of this biosynthetic pathway in *Capsicum annuum* chromoplasts. *Plant J* **8**: 417–424
- Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**: 13–34
- Ilag LL, Kumar AM, Soll D (1994) Light regulation of chlorophyll biosynthesis at the level of 5-aminolevulinic acid formation in *Arabidopsis*. *Plant Cell* **6**: 265–275
- Im CS, Beale SI (2000) Identification of possible signal transduction components mediating light induction of the *Gsa* gene for an early chlorophyll biosynthetic step in *Chlamydomonas reinhardtii*. *Planta* **210**: 999–1005
- Im CS, Grossman AR (2002) Identification and regulation of high light-induced genes in *Chlamydomonas reinhardtii*. *Plant J* **30**: 301–313
- Im CS, Matters GL, Beale SI (1996) Calcium and calmodulin are involved in blue light induction of the *gsa* gene for an early chlorophyll biosynthetic step in *Chlamydomonas*. *Plant Cell* **8**: 2245–2253
- Im CS, Zhang Z, Shrager J, Chang CW, Grossman AR (2003) Analysis of light and CO₂ regulation in *Chlamydomonas reinhardtii* using genome-wide approaches. *Photosynth Res* **75**: 111–125
- Janero DR, Barnett R (1982) Thylakoid membrane biogenesis in *Chlamydomonas reinhardtii* 137⁺. II. Cell cycle variation in the synthesis and assembly of pigments. *J Cell Biol* **93**: 411–416
- Johanningmeier U (1988) Possible control of transcript levels by chlorophyll precursors in *Chlamydomonas*. *Eur J Biochem* **177**: 417–424
- Johanningmeier U, Howell SH (1984) Regulation of light-harvesting chlorophyll-binding protein mRNA accumulation in *Chlamydomonas reinhardtii*. Possible involvement of chlorophyll synthesis precursors. *J Biol Chem* **259**: 13541–13549
- Jung KH, Hur J, Ryu CH, Choi Y, Chung YY, Miyao A, Hirochika H, An G (2003) Characterization of a rice chlorophyll-deficient mutant using the T-DNA gene-trap system. *Plant Cell Physiol* **44**: 463–472
- Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567–580
- Kropat J, Oster U, Rüdiger W, Beck CF (1997) Chlorophyll precursors are signals of chloroplast origin involved in light induction of nuclear heat-shock genes. *Proc Natl Acad Sci USA* **94**: 14168–14172
- Kruse E, Grimm B, Beator J, Kloppstech K (1997) Developmental and circadian control of the capacity for δ -aminolevulinic acid synthesis in green barley. *Planta* **202**: 235–241
- Kruse E, Mock HP, Grimm B (1995) Coproporphyrinogen III oxidase from barley and tobacco—sequence analysis and initial expression studies. *Planta* **196**: 796–803
- Kumar AM, Chaturvedi S, Söll D (1999) Selective inhibition of HEMA gene expression by photooxidation in *Arabidopsis thaliana*. *Phytochemistry* **51**: 847–850
- Lake V, Willows RD (2003) Rapid extraction of RNA and analysis of transcript levels in *Chlamydomonas reinhardtii* using real-time RT-PCR: magnesium chelatase *chlH*, *chlD* and *chlI* gene expression. *Photosynth Res* **77**: 69–76
- Lange BM, Ghassseman M (2003) Genome organization in *Arabidopsis thaliana*: a survey for genes involved in isoprenoid and chlorophyll metabolism. *Plant Mol Biol* **51**: 925–948
- Lange BM, Rujan T, Martin W, Croteau R (2000) Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proc Natl Acad Sci USA* **97**: 13172–13177
- Lermontova I, Kruse E, Mock HP, Grimm B (1997) Cloning and characterization of a plastidal and a mitochondrial isoform of tobacco protoporphyrinogen IX oxidase. *Proc Natl Acad Sci USA* **94**: 8895–8900
- Leustek T, Smith M, Murillo M, Singh DP, Smith AG, Woodcock SC, Awan SJ, Warren MJ (1997) Siroheme biosynthesis in higher plants. Analysis of an S-adenosyl-L-methionine-dependent uroporphyrinogen III methyltransferase from *Arabidopsis thaliana*. *J Biol Chem* **272**: 2744–2752
- Levine RP (1969) The analysis of photosynthesis using mutant strains of algae and higher plants. *Annu Rev Plant Physiol* **20**: 523–540

- Lichtenthaler HK (1999) The 1-deoxy-D-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants. *Annu Rev Plant Physiol Plant Mol Biol* **50**: 47–65
- Liu Y, Roof S, Ye Z, Barry C, van Tuinen A, Vrebalov J, Bowler C, Giovannoni J (2004) Manipulation of light signal transduction as a means of modifying fruit nutritional quality in tomato. *Proc Natl Acad Sci USA* **101**: 9897–9902
- Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**: 402–408
- Lotan T, Hirschberg J (1995) Cloning and expression in *Escherichia coli* of the gene encoding beta-C-4-oxygenase, that converts beta-carotene to the ketocarotenoid canthaxanthin in *Haematococcus pluvialis*. *FEBS Lett* **364**: 125–128
- Ma YZ, Holt NE, Li XP, Niyogi KK, Fleming GR (2003) Evidence for direct carotenoid involvement in the regulation of photosynthetic light harvesting. *Proc Natl Acad Sci USA* **100**: 4377–4382
- Martens JH, Barg H, Warren MJ, Jahn D (2002) Microbial production of vitamin B₁₂. *Appl Microbiol Biotechnol* **58**: 275–285
- Matsuzaki M, Misumi O, Shin IT, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa E, Yoshida Y, et al (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* **428**: 653–657
- Matters GL, Beale SI (1994) Structure and light-regulated expression of the *gsa* gene encoding the chlorophyll biosynthetic enzyme, glutamate 1-semialdehyde aminotransferase, in *Chlamydomonas reinhardtii*. *Plant Mol Biol* **24**: 617–629
- Matters GL, Beale SI (1995) Blue-light-regulated expression of genes for two early steps of chlorophyll biosynthesis in *Chlamydomonas reinhardtii*. *Plant Physiol* **109**: 471–479
- McCormac AC, Fischer A, Kumar AM, Soll D, Terry MJ (2001) Regulation of *HEMA1* expression by phytochrome and a plastid signal during etiolation in *Arabidopsis thaliana*. *Plant J* **25**: 549–561
- McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**: 404–405
- Misawa N, Truesdale MR, Sandmann G, Fraser PD, Bird C, Schuch W, Bramley PM (1994) Expression of a tomato cDNA coding for phytoene synthase in *Escherichia coli*, phytoene formation in vivo and in vitro, and functional analysis of the various truncated gene products. *J Biochem (Tokyo)* **116**: 980–985
- Mock HP, Trainotti L, Kruse E, Grimm B (1995) Isolation, sequencing and expression of cDNA sequences encoding uroporphyrinogen decarboxylase from tobacco and barley. *Plant Mol Biol* **28**: 245–256
- Moseley JL, Page MD, Alder NP, Eriksson M, Quinn J, Soto F, Theg SM, Hippler M, Merchant S (2002) Reciprocal expression of two candidate di-iron enzymes affecting photosystem I and light-harvesting complex accumulation. *Plant Cell* **14**: 673–688
- Moseley J, Quinn J, Eriksson M, Merchant S (2000) The *Crd1* gene encodes a putative di-iron enzyme required for photosystem I accumulation in copper deficiency and hypoxia in *Chlamydomonas reinhardtii*. *EMBO J* **21**: 2139–2151
- Nagata N, Satoh S, Tanaka R, Tanaka A (2004) Domain structures of chlorophyllide *a* oxygenase of green plants and *Prochlorothrix hollandica* in relation to catalytic functions. *Planta* **218**: 1019–1025
- Nakayama M, Masuda T, Sato N, Yamagata H, Bowler C, Ohta H, Shioi Y, Takamiya K (1995) Cloning, subcellular localization and expression of CHL1, a subunit of magnesium-chelatase in soybean. *Biochem Biophys Res Commun* **215**: 422–428
- Naya H, Romero H, Carels N, Zavala A, Musto H (2001) Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*. *FEBS Lett* **501**: 127–130
- Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**: 1–6
- Niyogi KK (1999) Photoprotection revisited: genetic and molecular approaches. *Annu Rev Plant Physiol Plant Mol Biol* **50**: 333–359
- Okada K, Saito T, Nakagawa T, Kawamukai M, Kamiya Y (2000) Five geranylgeranyl diphosphate synthases expressed in different organs are localized into three subcellular compartments in *Arabidopsis*. *Plant Physiol* **122**: 1045–1056
- Oster U, Tanaka R, Tanaka A, Rüdiger W (2000) Cloning and functional expression of the gene encoding the key enzyme for chlorophyll b biosynthesis (CAO) from *Arabidopsis thaliana*. *Plant J* **21**: 305–310
- Panek H, O'Brian MR (2002) A whole genome view of prokaryotic haem biosynthesis. *Microbiol* **148**: 2273–2282
- Papenbrock J, Pfündel E, Mock HP, Grimm B (2000) Decreased and increased expression of the subunit CHL I diminishes Mg chelatase activity and reduces chlorophyll synthesis in transgenic tobacco plants. *Plant J* **22**: 155–164
- Pogson BJ, Rissler HM (2000) Genetic manipulation of carotenoid biosynthesis and photoprotection. *Philos Trans R Soc Lond B Biol Sci* **355**: 1395–1403
- Quarmby LM (1994) Signal transduction in the sexual life of *Chlamydomonas*. *Plant Mol Biol* **26**: 1271–1287
- Querol J, Campos N, Imperial S, Boronat A, Rodriguez-Concepcion M (2002) Functional analysis of the *Arabidopsis thaliana* GCPE protein involved in plastid isoprenoid biosynthesis. *FEBS Lett* **514**: 343–346
- Quinn JM, Nakamoto SS, Merchant S (1999) Induction of coproporphyrinogen oxidase in *Chlamydomonas* chloroplasts occurs via transcriptional regulation of *CpxI* mediated by copper response elements and increased translation from a copper deficiency-specific form of the transcript. *J Biol Chem* **274**: 14444–14454
- Ravanel S, Block MA, Rippert P, Jabrin S, Curien G, Rébeillé F, Douce R (2004) Methionine metabolism in plants. Chloroplasts are autonomous for *de novo* methionine synthesis and can import S-adenosylmethionine from the cytosol. *J Biol Chem* **279**: 22548–22557
- Reinbothe S, Reinbothe C, Apel K, Lebedev N (1996) Evolution of chlorophyll biosynthesis—the challenge to survive photooxidation. *Cell* **86**: 703–705
- Rochaix JD (2002) *Chlamydomonas*, a model system for studying the assembly and dynamics of photosynthetic complexes. *FEBS Lett* **529**: 34–38
- Rodriguez-Concepcion M, Boronat A (2002) Elucidation of the methylerythritol phosphate pathway for isoprenoid biosynthesis in bacteria and plastids. A metabolic milestone achieved through genomics. *Plant Physiol* **130**: 1079–1089
- Rohmer M (2003) Mevalonate-independent methylerythritol phosphate pathway for isoprenoid biosynthesis. Elucidation and distribution. *Pure Appl Chem* **75**: 375–387
- Ronen G, Carmel-Goren L, Zamir D, Hirschberg J (2000) An alternative pathway to β -carotene formation in plant chromoplasts discovered by map-based cloning of beta and old-gold color mutations in tomato. *Proc Natl Acad Sci USA* **97**: 11102–11107
- Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In S Krawetz, S Misener, eds, *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp 365–386
- Rüdiger W, Klement H, Helfrich M, Tanaka A, Schoch S, Oster U (1999) Redox reactions in the last steps of chlorophyll biosynthesis. In JH Argyroudi-Akoyunoglou, H Senger, eds, *The Chloroplast: From Molecular Biology to Biotechnology*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 185–190
- Santana MA, Tan FC, Smith AG (2002) Molecular characterisation of coproporphyrinogen oxidase from *Glycine max* and *Arabidopsis thaliana*. *Plant Physiol Biochem* **40**: 289–298
- Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* **4**: 222–230
- Sharp PM, Tuohy TM, Mosurski KR (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14**: 5125–5143
- Shrager J, Hauser C, Chang CW, Harris EH, Davies J, McDermott J, Tamse R, Zhang Z, Grossman AR (2003) *Chlamydomonas reinhardtii* genome project. A guide to the generation and use of the cDNA information. *Plant Physiol* **131**: 401–408
- Siefermann-Harms D (1987) The light-harvesting and protective functions of carotenoids in photosynthetic membranes. *Physiol Plant* **69**: 561–568
- Simpson CL, Stern DB (2002) Mining the treasure trove of algal chloroplast genomes: surprises in architecture and gene content, and their functional implications. *Plant Physiol* **129**: 957–966
- Small I, Peeters N, Legeai F, Lurin C (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* **4**: 1581–1590
- Staden R (1996) The Staden sequence analysis package. *Mol Biotechnol* **5**: 233–241

- Steinbacher S, Kaiser J, Gerhardt S, Eisenreich W, Huber R, Bacher A, Rohdich F** (2003) Crystal structure of the type II isopentenyl diphosphate:dimethylallyl diphosphate isomerase from *Bacillus subtilis*. *J Mol Biol* **329**: 973–982
- Steinbrenner J, Linden H** (2001) Regulation of two carotenoid biosynthesis genes coding for phytoene synthase and carotenoid hydroxylase during stress-induced astaxanthin formation in the green alga *Haematococcus pluvialis*. *Plant Physiol* **125**: 810–817
- Strand A, Asami T, Alonso J, Ecker JR, Chory J** (2003) Chloroplast to nucleus communication triggered by accumulation of Mg-protoporphyrin IX. *Nature* **421**: 79–83
- Suzuki JY, Bollivar DW, Bauer CE** (1997) Genetic analysis of chlorophyll biosynthesis. *Annu Rev Genet* **31**: 61–89
- Tanaka A, Ito H, Tanaka R, Tanaka NK, Yoshida K, Okada K** (1998) Chlorophyll *a* oxygenase (CAO) is involved in chlorophyll *b* formation from chlorophyll *a*. *Proc Natl Acad Sci USA* **95**: 12719–12723
- Thompson JD, Higgins DG, Gibson TJ** (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680
- Tomitani A, Okada K, Miyashita H, Matthijs HC, Ohno T, Tanaka A** (1999) Chlorophyll *b* and phycobilins in the common ancestor of cyanobacteria and chloroplasts. *Nature* **400**: 159–162
- Tottey S, Block MA, Allen M, Westergren T, Albrieux C, Scheller HV, Merchant S, Jensen PE** (2003) *Arabidopsis* CHL27, located in both envelope and thylakoid membranes, is required for the synthesis of protochlorophyllide. *Proc Natl Acad Sci USA* **100**: 16119–16124
- Tracewell CA, Vrettos JS, Bautista JA, Frank HA, Brudvig GW** (2001) Carotenoid photooxidation in photosystem II. *Arch Biochem Biophys* **385**: 61–69
- Turkina MV, Villarejo A, Vener AV** (2004) The transit peptide of CP29 thylakoid protein in *Chlamydomonas reinhardtii* is not removed but undergoes acetylation and phosphorylation. *FEBS Lett* **564**: 104–108
- Ujwal ML, McCormac AC, Goulding A, Kumar AM, Soll D, Terry MJ** (2002) Divergent regulation of the HEMA gene family encoding glutamyl-tRNA reductase in *Arabidopsis thaliana*: expression of HEMA2 is regulated by sugars, but is independent of light and plastid signaling. *Plant Mol Biol* **50**: 83–91
- Van de Peer Y, De Wachter R** (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput Appl Biosci* **10**: 569–570
- Vavilin DV, Vermaas WFJ** (2002) Regulation of the tetrapyrrole biosynthetic pathway leading to heme and chlorophyll in plants and cyanobacteria. *Physiol Plant* **115**: 9–24
- von Heijne G, Steppuhn J, Herrmann RG** (1989) Domain structure of mitochondrial and chloroplast targeting peptides. *Eur J Biochem* **180**: 535–545
- Walker CJ, Willows RD** (1997) Mechanism and regulation of Mg-chelatase. *Biochem J* **327**: 321–333
- Walter MH, Hans J, Strack D** (2002) Two distantly related genes encoding 1-deoxy-d-xylulose 5-phosphate synthases: differential regulation in shoots and apocarotenoid-accumulating mycorrhizal roots. *Plant J* **31**: 243–254
- Woitsch S, Romer S** (2003) Expression of xanthophyll biosynthetic genes during light-dependent chloroplast differentiation. *Plant Physiol* **132**: 1508–1517
- Wright F** (1990) The effective number of codons used in a gene. *Gene* **87**: 23–29
- Xu H, Vavilin D, Vermaas W** (2001) Chlorophyll *b* can serve as the major pigment in functional photosystem II complexes of cyanobacteria. *Proc Natl Acad Sci USA* **98**: 14168–14173
- Yamamoto HY, Bassi R** (1996) Carotenoids: localization and function. In DR Ort, CF Yocum, eds, *Oxygenic Photosynthesis: The Light Reactions*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 539–563
- Yamamoto HY, Bugos RC, Hieber AD** (1999) Biochemistry and molecular biology of the xanthophyll cycle. In HA Frank, AJ Young, G Britton, RJ Cogdell, eds, *The Photochemistry of Carotenoids*, Vol 8. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 293–303
- Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al** (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**: 79–92
- Zhang Z, Shrager J, Chang C-W, Vallon O, Grossman AR** (2004) Genome based analysis of sulfur deprivation of wild-type cells and the *sac1* mutant of *Chlamydomonas*. *Eukaryot Cell* **3**: 1331–1348
- Zhu XF, Suzuki K, Saito T, Okada K, Tanaka K, Nakagawa T, Matsuda H, Kawamukai M** (1997) Geranylgeranyl pyrophosphate synthase encoded by the newly isolated gene GGPS6 from *Arabidopsis thaliana* is localized in mitochondria. *Plant Mol Biol* **35**: 331–341