Protist, Vol. 155, 65–78, March 2004
http://www.elsevier-deutschland.de/protist
Published online 19 March 2004

Protist

ORIGINAL PAPER

Dinoflagellate Expressed Sequence Tag Data Indicate Massive Transfer of Chloroplast Genes to the Nuclear Genome

Tsvetan R. Bachvaroff[a], Gregory T. Concepcion[a], Carolyn R. Rogers[a], Eliot M. Herman[b], and Charles F. Delwiche[a,1]

[a] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA
[b] Plant Genetics Research Unit, USDA/ARS, Donald Danforth Plant Science Center, 975 North Warson Rd, St. Louis, MO 63132, USA

Submitted July 7, 2003; Accepted October 15, 2003
Monitoring Editor: Michael Melkonian

The peridinin-pigmented plastids of dinoflagellates are very poorly understood, in part because of the paucity of molecular data available from these endosymbiotic organelles. To identify additional gene sequences that would carry information about the biology of the peridinin-type dinoflagellate plastid and its evolutionary history, an analysis was undertaken of arbitrarily selected sequences from cDNA libraries constructed from *Lingulodinium polyedrum* (1012 non-redundant sequences) and *Amphidinium carterae* (2143). Among the two libraries 118 unique plastid-associated sequences were identified, including 30 (most from *A. carterae*) that are encoded in the plastid genome of the red alga *Porphyra*. These sequences probably represent *bona fide* nuclear genes, and suggest that there has been massive transfer of genes from the plastid to the nuclear genome in dinoflagellates. These data support the hypothesis that the peridinin-type plastid has a minimal genome, and provide data that contradict the hypothesis that there is an unidentified canonical genome in the peridinin-type plastid. Sequences were also identified that were probably transferred directly from the nuclear genome of the red algal endosymbiont, as well as others that are distinctive to the Alveolata. A preliminary report of these data was presented at the Botany 2002 meeting in Madison, WI.

Introduction

Dinoflagellates are environmentally and economically important flagellates that are common in both freshwater and marine environments. About half of all dinoflagellates are photosynthetic. As do all photosynthetic eukaryotes, dinoflagellates rely on a plastid, an endosymbiotic organelle derived from a previously free-living cyanobacterium, to perform photosynthesis. Although fundamentally similar to the chloroplasts of plants and algae – and derived from a common ancestor – the plastids of dinoflagellates have a number of unique characteristics (Delwiche et al. 2003). The majority of photosynthetic dinoflagellates rely on a distinctive peridinin-containing plastid, but a number of other plastid types are found within the group, apparently the result of several independent symbiotic events (Delwiche 1999). The typical, peridinin-type plastid is pigmented with

[1] Corresponding author;
fax 1 301-314-9082
e-mail delwiche@umd.edu

1434-4610/04/155/01-065 $ 30.00/0

chlorophylls *a*-, *c*- and peridinin, is surrounded by three unit membranes, and has thylakoids stacked in groups of three (van den Hoek et al. 1995). Among the distinctive properties of the peridinin-type plastid are a chloroplast genome that is thought to consist entirely of single-gene minicircles (Barbrook and Howe 2000; Hiller 2001; Zhang et al. 1999), a water soluble light harvesting complex composed of a chlorophyll *a*-/*c*- and peridinin binding protein, and reliance upon a nuclear-encoded form II rubisco of a type known elsewhere only from anoxygenic photosynthetic bacteria (Morse et al. 1995; Rowan et al. 1996).

The dinoflagellate host cell is similarly distinctive, and many dinoflagellates can easily be recognized by their flagellar arrangement, thecal plates, and conspicuous nucleus with permanently condensed chromosomes (Graham and Wilcox 2000; van den Hoek et al. 1995). There are no recognizable histones or nucleosomes, and the nuclear genome is very large (10^{10}–10^{12} bp, i.e., up to 100-fold larger than the human genome; Rizzo 1987; Rizzo and Noodén 1972). These unusual features led some authors to view the dinoflagellate nucleus as an outgroup to other eukaryotes, and its organization has sometimes referred to as "mesokaryotic" or "dinokaryotic" to emphasize its uniqueness (Dodge and Greuet 1987). However, ultrastructural and molecular phylogenetic studies unequivocally place dinoflagellates with ciliates and apicomplexans in a monophyletic group known as the Alveolata (Cavalier-Smith 1993; Gajadhar et al. 1991; Wolters 1991).

Consequently, the plastids of dinoflagellates are important not only for their photosynthetic function in a key phytoplankton group that retains the ability to acquire endosymbiotic organelles. The acquisition of organelles is intruiging particularly in view of the complex interactions between organellar and nuclear genome.

To study the incorporation of the peridinin-type plastid in the dinoflagellate cell, we undertook an expressed sequence tag (EST) survey of two peridinin containing dinoflagellates as an inexpensive alternative to whole-genome sequencing in a case where the genome is extremely large (Adams et al. 1991). The results are striking, and indicate that many typically plastid-encoded genes are encoded in the nuclear genome in dinoflagellates. Transfer seems to have occurred from both the plastid and the (red algal) intermediate chloroplast host. This survey has also identified genes that appear to be shared only by dinoflagellates and *Plasmodium*. These data can provide insight into the basic biology of dinoflagellates, the processes governing plastid acquisition, and the evolution of Alveolates.

Results

Overview

A total of 4899 ESTs were determined from the two cDNA libraries, 1519 from *Lingulodinium polyedrum* (Stein) Dodge 1989, strain 70 *(= Gonyaulax polyedra* GenBank accessions CD809360–CD810879*)*, and 3380 from *Amphidinium carterae* Hulburt 1957, CCMP 1314 (GenBank accessions CF064497–CF067877). Both libraries were unidirectional, and most reads were from the 5′ end. Sequencing of the *L. polyedrum* library, which was not constructed in house, commenced while the *A. carterae* library was being prepared. The reads from the *L. polyedrum* library had an average length of 506 bp, of which those with a bit score above 100 had an average length of 583. Sequencing on the *L. polyedrum* library was halted when the *A. carterae* library was ready for sequencing. The most abundant transcript from the *L. polyedrum* library was the peridinin-chlorophyll binding protein, which constituted 45 out of 1519 clones, or 3%. A total of 193 gene sequences were found more than once, accounting for 709 of 1519 sequences, or 46.7%, of all ESTs. There were 819 singletons (i.e., sequences found only once). To measure cumulative error during library amplification and sequencing 10,435 bp of sequence from the 34 different sequencing reads of the apparently invariant peridinin-chlorophyll binding protein were compared. These analyses indicate a maximum error rate in the first 350 bases of less than 0.05%. The average insert size for this library was quite low, but only clones with an apparent size of >500 bp were selected for sequencing. When sequencing on the *L. polyedrum* library was halted, the last plate had over 62% novel sequences, suggesting that this library was far from exhaustion.

The modified vector used for the *A. carterae* library permitted a somewhat longer read than for *L. polyedrum*, and the average read length for the 3380 clones sequenced was 650 bp. The average insert size based on EcoRI and PstI digests of the initial 192 clones was 1.9 kb. The error rate for *A. carterae* was calculated from 9,845 bp of redundant reads from 9 clones, and was 0.05%. Blast analysis identified 1347 sequences with a bit score above 50 (with 609 > 100). As would be expected, and consistent with the results from *L. polyedrum*, longer sequences were more likely to be identified by blast; those with a bit score above 50 had an average length of 688, and those above 100 of 703 bp. In the *A. carterae* library the two most abundant transcripts were EF-1α and an unidentified sequence with partial similarity to a viral protein, each of which constituted less than 1% of the clones.

Redundant ESTs and those from closely related gene families were clustered with Sequencher (GeneCodes, Ann Arbor MI), which uses a modified Smith-Waterman algorithm to find the globally optimal alignment of sequences that meet minimum overlap criteria (40 bp, 70% identity). After clustering the *L. polyedrum* library had 1012 non-redundant sequences (i.e., unique entities), several of which may represent nonoverlapping reads from equivalent ESTs. Where practical, independent reads that appeared to be from the same transcript were grouped, but this is not feasible in cases where no homolog is known and no overlap was found, so the probable number of proteins represented by these data is less than 1012. Similar analyses were performed for the *A. carterae* data. Of the 3380 ESTs from *A. carterae,* 1702 were grouped into 621 clusters, leaving 1522 singletons and a total of 2143 non-redundant sequences. Databases presenting the *L. polyedrum* and *A. carterae* EST data are available at http://oxrid.umd.edu, and the data have been deposited in GenBank.

Plastid-Associated Sequences

Initial identification of likely plastid-associated sequences (defined here as sequences that are expressed in or evolutionarily derived from the plastid) was performed by blast analysis. ESTs were considered likely to be plastid-targeted if blast analysis identified them as homologous to cyanobacterial or plastid gene sequences. Based on blastx scores and clustering, 38 plastid-associated genes were identified in the *L. polyedrum* library. Of these, 4 are known to be plastid-encoded in *Porphyra*. In the *A. carterae* library 99 plastid-associated genes were identified, including 27 that are plastid-encoded in *Porphyra*. Clustering and elimination of redundancy between

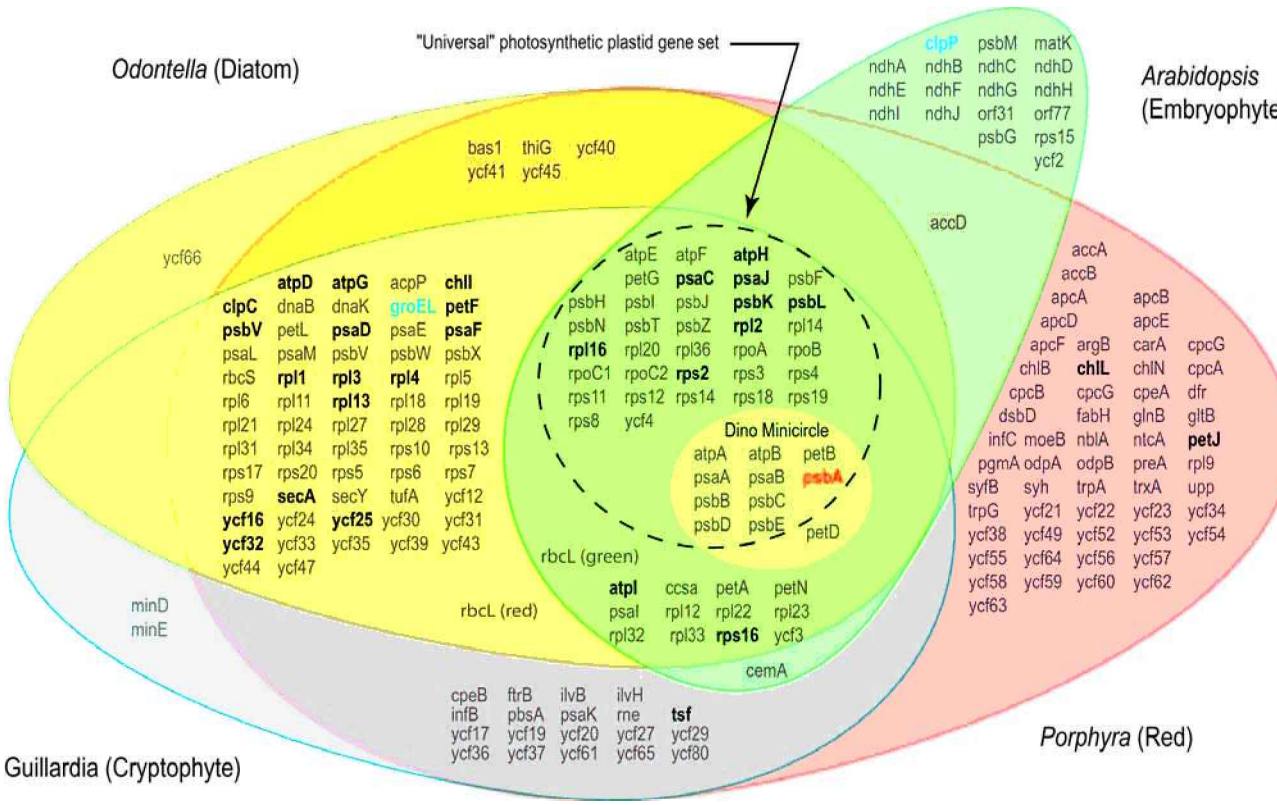

Figure 1. Venn diagram comparing the gene contents of the plastid genomes of *Porphyra purpurea* (Rhodophyta), *Odontella sinensis* (Bacillariophyceae), *Guillardia theta* (Cryptophyta), *Arabidopsis thaliana* (plant), and known peridinin-type plastid minicircles. The dotted line indicates those genes that are found in these genomes as well as in *Cyanophora paradoxa*, *Nephroselmis olivacea*, *Euglena gracilis, Zea mays, Oryza sativa, Nicotiana tabacum, Spinacia oleracea, Oenothera alata, Pinus thunbergii, Marchantia polymorpha, Mesostigma viride, Cyanidium caldarium.* Genes shown in boldface have been identified from the cDNA data presented here. Blue indicates possible homology to nuclear-encoded copies of the gene. Red indicates a minicircle-encoded gene found in the EST project. Additional comparisons are presented by Grzebyk (Grzebyk et al. 2003) and Martin et al. (Martin et al. 2002). Gene nomenclature follows Martin et al. (Martin et al. 2002).

the two libraries produced a non-redundant set of 118 candidate plastid-associated sequences. Of these, 30 genes – most of which were identified from the *A. carterae* library – are encoded in the plastid genome of *Porphyra* (Table 2; Reith and Munholland 1995). The remainder is presumed to be nuclear-encoded in *Porphyra* and most other taxa (Table 2), although in many cases the location and presence of the gene has not been well characterized. These data are summarized and compared to the plastid genome content of other species in Figure 1.

The ESTs that represented genes that are encoded in the plastid genome in *Porphyra* (Table 1) were fully sequenced to verify the presence of poly-A tails and to provide full-length sequences for analysis. Among these, some cDNAs that encode the same gene were found to have substantial sequence variation. For example, cDNAs encoding *atpH* were found 10 times from *L. polyedrum*, and these sequences formed five distinct clusters. The sequences assembled into a single, 452 nucleotide transcript, consisting of a 249 base "mature protein"

Table 1. Dinoflagellate ESTs present in the *Porphyra* plastid genome, sorted by bitscore.

gene[a]	bit-score[b]	e value[c]	clone reference[d]	variation[e]	polyA[f]	SignalP[g]	ChloroP[h]	Source[i]	Accession
chlI	358	1.0×10^{-119}	AcContig[0857]	family	yes	0.205	0.568*	A	CF067189
atpI	282	7.0×10^{-76}	AcContig[1157]	family	yes	0.817*	0.489	A	CF065976
chlL	234	1.0×10^{-128}	AcContig[0737]	1	yes	0.395	0.559*	A	CF064591
ycf16	187	3.0×10^{-75}	AcContig[1099]	1	yes	0.212	0.494	A	CF064637
rps2	176	1.0×10^{-42}	AcContig[0749]	1	yes	0.093	0.451	A	CF064824
petK	160	4.0×10^{-37}	AcContig[0964]	family	yes	0.932*	0.548*	A	CF066266
petF	152	3.0×10^{-36}	AcContig[1605]	family	yes	0.589*	0.486	Both	CF067664
psaD	148	1.0×10^{-34}	AcContig[0733]	1	yes	0.836*	0.571*	A	CF064527
rpl1	139	9.0×10^{-32}	AcContig[0762]	1	yes	0.247	0.471	A	CF064976
rpl16	138	1.0×10^{-31}	Ac1119	–	yes	0.743*	0.532*	A	CF064566
psaC	127	3.0×10^{-28}	AcContig[1109]	family	yes	0.823*	0.518*	A	CF066614
rpl13	114	1.0×10^{-24}	AcContig[1636]	1	yes	0.763*	0.441	A	CF066354
petJ	110	1.0×10^{-24}	Ac5812	family	yes	0.355	0.487	A	CF067105
secA	108	6.0×10^{-41}	AcContig[1437]	1	yes	N.A.	N.A.	A	CF066408
psaF	103	1.0×10^{-23}	Ac977	–	yes	0.290	0.449	A	CF067650
rpl3	97	5.0×10^{-23}	AcContig[1546]	1	yes	0.736*	0.552*	A	CF067587
psaE	87	1.0×10^{-16}	Ac6843	–	yes	0.567*	0.481	A	CF067821
ftsH	85	7.0×10^{-16}	Ac1454r	–	no	N.A.	N.A.	Both	CF064829
atpH	84	9.0×10^{-16}	AcContig[0805]	family	yes	0.879*	0.516*	Both	CF067275
tsf	81	3.0×10^{-32}	AcContig[1710]	1	no	0.040	0.427	A	CF067081
atpG	77	2.0×10^{-13}	Ac1899	–	yes	0.580*	0.494	A	CF065024
rpl4	75	6.0×10^{-19}	AcContig[1547]	1	yes	0.255	0.482	A	CF066238
clpC	69	1.0×10^{-22}	AcContig[1539]	1	no	N.A.	N.A.	A	CF065755
rps1	69	1.0×10^{-17}	AcContig[1662]	1	yes	0.379	0.445	A	CF065490
atpD	64	1.0×10^{-09}	Lp587	–	yes	N.A.	N.A.	L	CD810773
rpl33	63	3.0×10^{-09}	Ac6830	–	yes	0.725*	0.555*	A	CF067798
psaJ	41	0.007	Ac1256	–	yes	0.226	0.430	A	CF064650
psbY	38	0.046	Ac6675	–	yes	0.319	0.509*	A	CF067444
psbL	35	0.17	Ac6375	–	yes	0.699*	0.541*	A	CF067332
psbK	33	0.73	AcContig[1306]	1	yes	0.396	0.436	A	CF066016

[a] Gene name following Martin et al., 2002.
[b] Highest bitscore in blastx analysis.
[c] Corresponding e-value from blastx analysis.
[d] Best hit identifier in the dinoflagellate EST database.
[e] Number of sequence types in multiply sampled ESTs, dash indicates unique EST.
[f] Presence of poly-A tail.
[g] SignalP mean S score; * indicates values that are significant (>0.5).
[h] ChloroP score; * indicates values that are significant (>0.5).
[i] Source L = *Lingulodinium polyedrum* A = *Amphidinium carterae.*

that corresponded well with homologous sequences from several plastid genomes, and a 204 base 5′ extension that encodes a candidate targeting peptide. However, despite agreement among these sequences on overall gene structure, there were numerous point mutations among the five clusters (within-cluster sequences were identical). Considering just the 249 bases of the putative mature protein, the most divergent pair of clusters Lp3266 (CD810707) *vs.* Lp102 (CD810870) had 33 nucleotide substitutions (13%), 31 of which were in third codon-position. The 5′ leader sequence was present in all clusters, and showed as many as 53 total substitutions in 204 bases (26%), 34 of which were in third codon-position. The amino acid translations and hydropathy plots of two different leader sequences for the *atpH* gene in *L. polyedrum* and *A. carterae* are shown in Figure 2. A similar pattern of differences in transit peptides was found for the genes *psbO* (Fig. 3) and *psaC* (data not shown), where greater variation was present in the leader than in the mature protein.

There was little contamination of the library with minicircle gene products. In the entire survey, only two sequences were identified that correspond to genes that have been identified on single-gene minicircles. One of these, Ac3135 (CF065874), is a perfect match to the published *A. operculatum psbA* minicircle sequence (Barbrook and Howe 2000) and consequently seems likely to be a genuine minicircle gene contaminating the poly-A fraction. The other is not a perfect match to any published sequence, but has a best blastn hit to the *Heterocapsa triquetra* plastid LSU rRNA sequence.

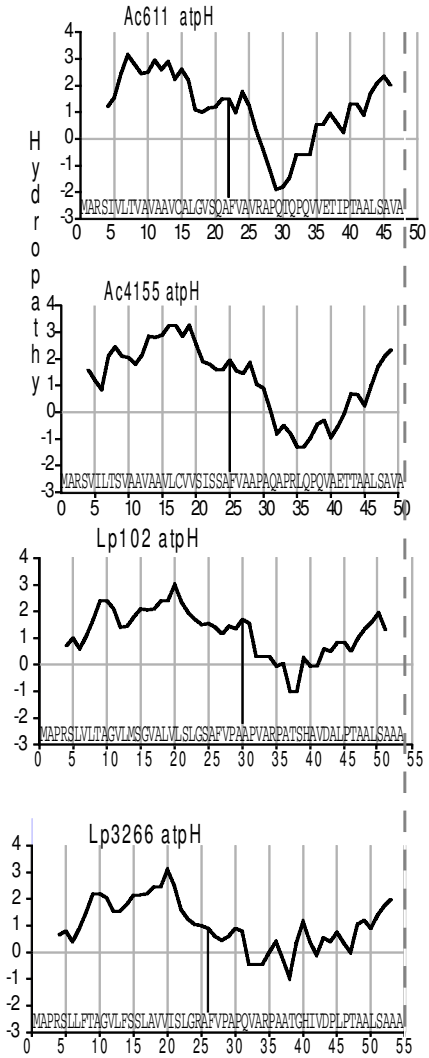

Figure 2. Putative transit peptides from gene products of different *atpH loci*. Kyle-Doolittle hydropathy plots are shown with a window size of 7 amino acids. The chloroplast cleavage sites were inferred from an alignment of mature proteins. The predicted signal sequence cleavage sites are indicated with a vertical line.

Signal Peptides

Some of these ESTs had leader sequences that were consistent with published descriptions of transit peptides in secondary plastids where the proteins are initially targeted to the ER (Ishida et al.

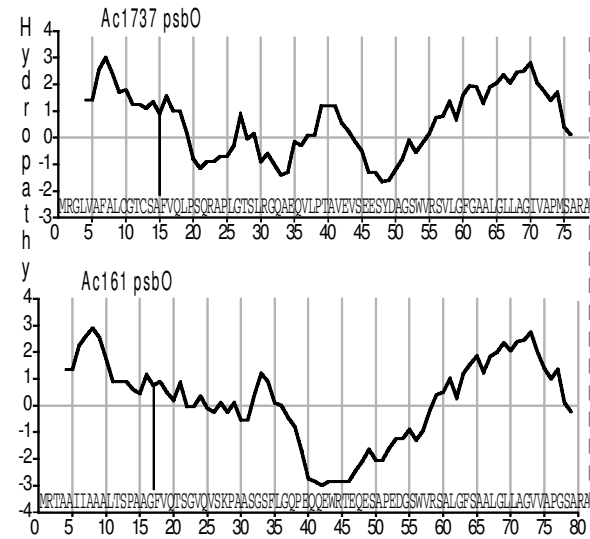

Figure 3. Putative transit peptides from different copies of the *psbO* gene product. Kyle-Doolittle hydropathy plots are shown with a window size of 7 amino acids. The cleavage sites were inferred from a multiple sequence alignment. The chloroplast cleavage sites were inferred from an alignment of mature proteins. The predicted signal sequence cleavage sites are indicated with a vertical line.

Table 2. Plastid-associated genes not in *Porphyra* plastid genome (putative nuclear-to-nuclear transfers), sorted by bitscore.

Gene Name	length[a]	bitscore[b]	e value[c]	Clone reference[d]	SignalP[e]	ChloroP[f]	Accession[g]
Ribulose bisphosphate carboxylase-oxygenase	1166	711	0.0	LpContig[0435]	0.060	0.429	CD810786
Glutamate semialdehyde synthase	1585	622	1.0×10^{-177}	AcContig[0832]	0.714*	0.515*	CF067231
Phosphoenol pyruvate synthase	1329	606	1.0×10^{-172}	LpContig[0475]	N.A.	N.A.	CD810119
Glyceraldehyde 3-phosphate dehydrogenase	1207	590	1.0×10^{-167}	LpContig[0425]	0.900*	0.565*	CD810603
Fructose-1,6-bisphosphate aldolase classII	1481	571	1.0×10^{-161}	AcContig[1111]	0.271	0.525*	CF067629
Peridinin chlorophyll protein (PCP)	1326	560	1.0×10^{-158}	LpContig[0334]	0.351	0.487	CD809573
Light harvesting complex (LHC)	1212	541	1.0×10^{-153}	AcContig[0799]	0.744*	0.571*	CF066495
Histidine-tRNA ligase archeal	1438	406	1.0×10^{-112}	AcContig[1033]	0.051	0.450	CF064869
Oxygen evolving enhancer protein (*psbO*)	1182	399	1.0×10^{-110}	AcContig[0758]	0.878*	0.513*	CF067369
Coproporphyrinogen III oxidase	1246	360	3.0×10^{-98}	AcContig[0734]	0.169	0.552*	CF064552
Porphobilinogen synthase	1480	344	7.0×10^{-97}	AcContig[1562]	0.640*	0.557*	CF066269
Transketolase 1 chloroplast	720	335	3.0×10^{-91}	Ac1168	N.A.		CF064604
Uroporphyrinogen decarboxylase (*uroD*)	1455	313	4.0×10^{-84}	AcContig[0828]	0.269	0.558*	CF066269
Malonyl CoA:ACP transacyl carrier (*fabD*)	1306	297	2.0×10^{-79}	AcContig[0959]	0.381	0.586*	CF067271
Aconitate hydrase	556	292	2.0×10^{-78}	Lp146b	N.A.		CD809560
Violaxanthin de-epoxidase precursor	1498	284	2.0×10^{-75}	AcContig[1564]	0.902*	0.482	CF066890
Ferredoxin NADP reductase	758	272	9.0×10^{-72}	AcContig[1305]	0.174	0.557*	CF067646
Mg protoporphyrin methyltransferase (*chlM*)	1030	263	3.0×10^{-69}	AcContig[0790]	0.288	0.483	CF066233
Phosphoribulokinase	1281	256	1.0×10^{-73}	AcContig[0779]	0.101	0.438	CF065476
Triosephosphateisomerase	1077	253	2.0×10^{-69}	AcContig[1664]	0.070	0.443	CF066220
Phosphoserine aminotransferase	771	250	3.0×10^{-67}	Ac5574	0.157	0.438	CF066962
Inorganic pyrophosphatase	836	248	1.0×10^{-64}	Ac4379	0.237	0.525*	CF066545
Mg chelatase subunit (*chlD*)	805	245	2.0×10^{-63}	AcContig[0766]	0.566*	0.522*	CF066220
Putative nucleotide-sugar dehydratase	728	229	7.0×10^{-64}	Lp1334	0.743*	0.460	CD809551
Flavoprotein cyanobacterial hits only	704	208	5.0×10^{-53}	Lp4457	N.A.		CD810473
Hydroxymethylbilane synthase	1272	207	1.0×10^{-76}	AcContig[1677]	0.263	0.480	CF065576
UDPglucose-starch glucosyltransferase	1305	204	3.0×10^{-51}	AcContig[1286]	0.241	0.599*	CF065409
Starch phosphorylase H	609	204	6.0×10^{-52}	Lp1271	N.A.		CD809492
Iron superoxide dismutase	579	201	7.0×10^{-51}	Lp2142	0.133	0.431	CD809812
Plastid mRNA binding protein	771	200	2.0×10^{-50}	Ac1315ra	N.A.		CF064730
Fructose 1,6-bisphosphatase	776	187	2.0×10^{-46}	Lp1187r	N.A.		CD810868
2–oxoglutarate/malate translocator	757	181	8.0×10^{-45}	Ac580r	N.A.		CF067093
ATP synthase gamma chain (*atpC*)	1000	180	3.0×10^{-44}	AcContig[0867]	0.705*	0.536*	CF065394
3-oxoacyl-(acyl-carrier-protein) (*fabB*)	604	169	3.0×10^{-41}	AcContig[1342]	N.A.		CF065186
[h]GI:16125539 Caulobacter crescentus	717	164	1.0×10^{-39}	LpContig[0273]	0.417	0.474	CD810585
Glutamine syntase III	845	162	5.0×10^{-39}	AcContig[1175]	0.133	0.455	CF065585
Thioredoxin reductase	739	160	2.0×10^{-38}	AcContig[1400]	0.249	0.470	CF064926
Protease (*clpP1*)	685	159	3.0×10^{-38}	Ac3109	0.217	0.537*	CF065849
Glutamyl tRNA synthetase cytosolic?	430	159	1.0×10^{-38}	Ac6685	N.A.		CF067450
Putative nitrate transporter	651	154	1.0×10^{-36}	Lp43	N.A.		CD810421
Methyltransferase	1245	153	7.0×10^{-36}	AcContig[1313]	0.048	0.464	CF066859
[h]GI:22987108 Burkholderia fungorum	607	151	7.0×10^{-36}	Lp42	N.A.		CD810372
Alanine aminotransferase	771	139	6.0×10^{-32}	AcContig[1243]	N.A.		CF066050
Farnesylpyrophosphate synthase	694	139	3.0×10^{-32}	LpContig[0522]	0.062	0.433	CD810787

Phosphoglycolate phosphatase	739	137	1.0×10^{-31}	AcContig[1076]	0.408	0.498	CF066481
Aspartyl protease? Chloroplast nucleoid binding?	649	136	2.0×10^{-31}	Ac4923	N.A.		CF066716
Isocitrate lyase	659	134	1.0×10^{-30}	AcContig[1372]	N.A.		CF065691
Pyrophosphatase	499	128	3.0×10^{-29}	Lp374a	0.158	0.468	CD810255
Putative CP membrane-associated 30 kD protein	582	127	8.0×10^{-29}	Ac6374	0.573*	0.531*	CF067331
cGI:27382321 Bradyrhizobium japonicum	726	125	8.0×10^{-28}	Ac2672r	N.A.		CF065545
Glutathione peroxidase	692	125	7.0×10^{-28}	Ac3739	0.065	0.435	CF066302
Cytochrome B6-F complex iron-sulfur subunit (*petC*)	646	124	9.0×10^{-28}	AcContig[1301]	0.913*	0.549*	CF067078
[h]GI:23039345 Trichodesmium erythraeum	1403	119	2.0×10^{-25}	AcContig[0819]	0.518*	0.456	CF065847
Ketothiolase	544	113	2.0×10^{-25}	Ac6987	N.A.		CF067524
Monodehydroascorbate reductase	913	111	1.0×10^{-23}	Ac1976r	N.A.		CF065064
[h]GI:16330484 Synechocystis PCC 6803	699	110	1.0×10^{-23}	Ac3932	N.A.		CF066369
Photosystem II protein psbU	617	101	9.0×10^{-21}	Ac4283	0.857*	0.478	CF066491
Cobalbumin synthase cGI:17229243	730	99	8.0×10^{-20}	Ac1263r	N.A.		CF064660
Chaperone (*dnaJ hsp40*)	552	97	2.0×10^{-20}	Ac1889	N.A.		CF065020
[h]GI:17979159 Arabidopsis	829	97	3.0×10^{-19}	Ac5807	0.518*		CF067099
Elongation Factor G	488	95	4.0×10^{-19}	Ac2516	N.A.		CF065438
Phosphoglycerate mutase (*gpmB*)	831	92	7.0×10^{-18}	Ac5805	N.A.		CF067097
[h]GI:15242446 Arabidopsis	493	91	9.0×10^{-18}	Lp4069	N.A.		CD810334
Chaperone (*cpn60 groEL*)	516	90	1.0×10^{-17}	Ac6963	0.348	0.486	CF067852
Carbonic anhydrase	641	90	3.0×10^{-17}	Lp1702	N.A.		CD809667
PEP/phosphate translocator-like protein	460	86	2.0×10^{-16}	Ac5934	N.A.		CF067200
FKBP-type peptidyl-prolyl cis-trans isomerase	858	85	1.0×10^{-15}	AcContig[0742]	0.555*	0.547*	CF067531
Acyl-CoA dehydrogenase (*fadE2*)	607	82	4.0×10^{-15}	Ac6379	N.A.		CF067335
Phenazine biosynthesis protein	593	80	1.0×10^{-19}	Ac3034	N.A.		CF065787
Pyridoxamine 5-phosphate oxidase	747	79	4.0×10^{-17}	Ac2510	0.085	0.436	CF065431
[h]GI:16329601 Synechocystis sp. PCC 6803	567	79	5.0×10^{-14}	AcContig[1192]	N.A.		CF066047
Some dnaJ similarity + ferredoxin	1167	77	7.0×10^{-13}	AcContig[0840]	0.326	0.531*	CF064949
Ferredoxin component	739	73	4.0×10^{-12}	AcContig[0864]	0.782*	0.552*	CF067564
Thioredoxin	796	71	2.0×10^{-11}	Ac1329r	0.764*	0.468	CF064745
Peroxisome/chloroplast ascorbate peroxidase	1665	69	2.0×10^{-10}	AcContig[0879]	0.149	0.456	CF067296
[h]GI:18405058 Arabidopsis	702	67	2.0×10^{-10}	AcContig[0923]	N.A.		CF066120
[h]GI:22326972 Arabidopsis	767	67	4.0×10^{-10}	LpContig[0295]	0.147	0.531*	CD809939
Putative methionyl-tRNA synthetase	666	65	7.0×10^{-10}	Lp1215	0.202	0.508*	CD809452
[h]GI:22971207 ABC transporter Chloroflexus	747	63	5.0×10^{-09}	Ac4912	0.420	0.496	CF066709
[h]GI:17230532 Nostoc sp. PCC 7120	451	62	1.0×10^{-09}	Ac1473	0.197	0.495	CF064845
Photosystem II 11kd protein	781	57	2.0×10^{-07}	AcContig[0903]	0.435	0.551*	CF065401
ABC-type transport protein	592	57	2.0×10^{-07}	Lp1325	N.A.		CD809542
WD domain	538	55	7.0×10^{-07}	Ac3924	N.A.		CF066361
RNA-binding protein (*cp33*)	378	53	1.0×10^{-06}	Ac2021	N.A.		CF065099
[h]GI:13812240 Guillardia nucleomorph	785	53	6.0×10^{-06}	Ac4855	N.A.		CF066666
[h]GI:16329535 Synechocystis sp. PCC 6803	825	52	1.0×10^{-05}	Ac1749	0.339	0.436	CF064947
Chloroplast 28 kDa ribonucleoprotein	589	46	2.0×10^{-04}	Lp3507	N.A.		CD810170
Chloroplast 30 kDa ribonucleoprotein	610	45	0.001	LpContig[0480]	N.A.		CD810304

[a]Length of largest assembly or EST, in nucleotide bases. [b]Best blastx bitscore. [c]Best blastx e-value. [d]Best hit identifier in the dinoflagellate EST database. [e]SignalP mean S score; * indicates values that are significant (>0.5). [f]ChloroP score; * indicates values that are significant (>0.5). [g]GenBank accession number for the best scoring EST. [h]Indicates best hit to putative protein of unknown or unverified identity; the NCBI GI ("GenInfo") number is provided as a uniform identifier.

2000; Nassoury et al. 2003; Peltier et al. 2000; Schein et al. 2001; Zuegge et al. 2001). Signal peptides were detected in a greater proportion of proteins destined for the thylakoid membrane (8 out of 12 in Table 1), than in non-thylakoid proteins (5 out of 15 in Table 1), but exceptions occurred even when apparently full length sequences were found (i.e. *psaF* in figure 4). None of the targeting-prediction software tested consistently recognized all these leader sequences as targeting peptides (Tables 1, 2).

Nucleus to Nucleus Gene Transfer

Among the nuclear-encoded, plastid-targeted ESTs, the light harvesting complex (LHC) gene family stood out. There was a high diversity of LHC sequences, with 47 individual ESTs clustered into 20 nonredundant sequences in *A. carterae*. There was sequence variation within the nonredundant clusters, and only four of these consisted entirely of identical sequences. Similarly, in *L. polyedrum* 21 ESTs clustered into 14 nonredundant sequences, none of which was composed of identical reads. Several sequences had previously been reported from *A. carterae*, including four that form a single polyprotein array (Hiller et al. 1995). The presence of polyproteins was confirmed for *A. carterae*, with ESTs identified that seem to correspond to each of the four repeats and trans-repeat regions. Evidence of a homologous polyprotein array consisting of at least three repeats was found in *L. polyedrum*. The EST data included the previously identified sequences along with considerable additional diversity of LHC sequences in both *A. carterae* and *L. polyedrum*. Blast analyses placed nine of the nonredundant *A. carterae* sequences with previously known *A. carterae* sequences and eleven with LHCs from other organisms (including *Galdieria*, *Guillardia*, and *Vaucheria*). For *L. polyedrum*, eight nonredundant sequences clustered with the *A. carterae* sequences in blast analysis, while six clustered with sequences from other taxa.

Comparison to Plasmodium

The tags for which the top *Plasmodium* hit was also one of the top ten hits in unconstrained searches of the nonredundant database were examined in detail. Among these sequences were several that may be specific to the alveolates, i.e., they have relatively high blastx scores compared to *Plasmodium* and poor scores to anything else. For example the tag Ac5698 (CF067023) has a blastx bitscore of 221 (e value = 6.0×10^{-57}) to a hypothetical ORF from *Plasmodium*, GI:16805161, but no other significant hit in the nr database. Similarly the tag Lp1707 (CD809670) has a bitscore of 120 (e value = 3.0×10^{-26}) to hypothetical *Plasmodium* ORF, GI:23482968, while the next highest hit has a bitscore of 34 and an e-value of 2.5 (i.e., no better than would be expected by chance). An additional two tags have hits only to *Plasmodium* among eukaryotes, with all other hits being to bacteria: one, Ac7147 (CF067672),

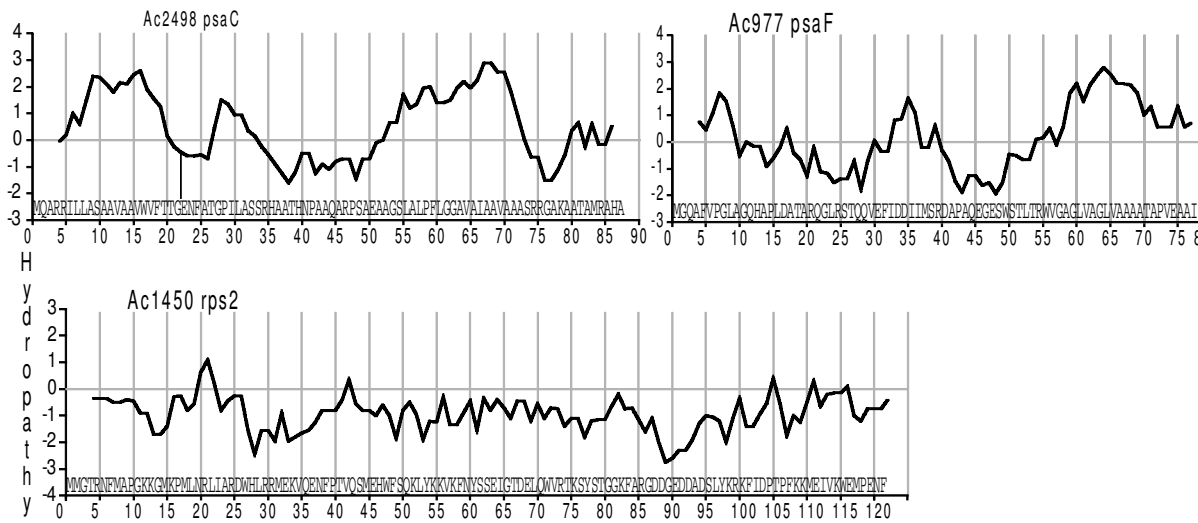

Figure 4. Putative transit peptides from the *psaC*, *psaF*, and *rps2* gene products. Kyle-Doolittle hydropathy plots are shown with a window size of 7 amino acids. The cleavage sites for these proteins were inferred from a multiple sequence alignment, although in the cases of the *psaF* and *rps2* gene products the cleavage sites are less certain. Only the *psaC* gene product contains a predicted signal sequence indicated with a vertical line.

apparently encodes a Leu/Phe aminoacyl-tRNA transferase, while the other, Ac1889 (CF065020), encodes a DNAJ-like chaperone. The latter sequence does show one relatively poor hit to *Arabidopsis*, suggesting the possibility that it is plastid-associated. Finally, two tags were both plastid and *Plasmodium* associated, but are not unique to the Alveolata: *fabD*, a malonyl CoA: ACP transacyl carrier, and *GcpE* (*IspG*) a gene involved in the DOXP pathway of isoprenoid biosynthesis (Hecht et al. 2001).

Discussion

Overview

This survey provides a suite of 4899 sequence tags representing roughly 3100 unique entities from two dinoflagellates, and these data can be used to understand gene transfer in peridinin dinoflagellates. The 1012 unique sequences from *L. polyedrum* and 2143 from *A. carterae* can be compared to 3267 unique sequences found in analysis of 10,154 ESTs from a normalized library from *Porphyra yezoensis* (Nikaido et al. 2000), which indicates that although the libraries were not explicitly normalized, they show high sequence diversity. Plastid-containing eukaryotes for which complete genome data are available include *Arabidopsis* with 25,500 genes (Arabidopsis Genome Initiative 2000) and *Plasmodium* with 5300 genes (Gardner et al. 2002). Both of these probably have somewhat streamlined genomes, but if one uses *Arabidopsis* as a base of comparison, the 2143 nonredundant sequences could account for as much as 8% of the genome complexity, and if the unicellular *Plasmodium* is a better basis for comparison this fraction could be substantially higher.

Evidence that the novel sequences presented here are encoded in the dinoflagellate nuclear genome includes poly-A tails, leader sequences, and the presence of a gene family for many genes. Because the nuclear location of the 30 genes that are encoded in the plastid genome of *Porphyra* is surprising and important to this study, these sequences were examined in detail. Clones were fully sequenced to verify the presence and terminal location of a poly-A tail, which was identified in all but three of the sequences (Table 1). In addition, 16 of these 31 sequences have a 5′ polypeptide extension that is scored by SignalP or ChloroP above 0.5, corresponding well to characterized targeting peptides. Of the 12 that were found more than once, 7 show sequence variation consistent with the presence of multiple alleles, a hallmark of nuclear-encoded genes (Table 1). Minicircle genes, although probably expressed at high levels, were essentially absent from the cDNA data.

The dinoflagellate cell is a potentially complex combination of several genomes. In addition to the nuclear and mitochondrial genomes of the host cell, there are possible genetic contributions from the plastid, mitochondrial, and nuclear genomes of the red alga that contributed the plastid. Careful sequence analysis is necessary to identify both the likely phylogenetic origin of the sequences and their probable compartmentalization in the cell. The sequences listed in Table 1 are homologous to plastid-encoded genes in *Porphyra*, and are almost certainly originally of plastid origin. Those in Table 2 are not in the *Porphyra* plastid genome, and information about localization and expression varies greatly depending upon the gene and organism in question.

Chloroplast to Nucleus Gene Transfer

A substantial number of the plastid-associated ESTs found in this study encode genes that are in the chloroplast genome in other organisms (Fig. 1). Because the peridinin-type plastid is thought to be ultimately derived from a red alga, the most appropriate comparison is to *Porphyra*, but a striking number of genes have been transferred even in comparison to the relatively depauperate plastid genomes of green algae and plants. Of the 31 genes found that are encoded in the chloroplast genome of *Porphyra* (Fig. 1, Table 1), eight are present in all known photosynthetic chloroplast genomes (Martin et al. 2002), and encode ribosomal proteins, ATP synthase, and photosystem components (Table 1). Given that these data represent an arbitrary subset of all of the plastid-associated genes in the nuclear genome, they suggest that in dinoflagellates the transfer of genes from the chloroplast to the nuclear genome has been more extensive than in any other group of organisms.

Two of the otherwise exclusively plastid encoded genes (*atpH* and *psaC*) exist in at least two alleles with distinctly different transit peptides. Transit peptides for these genes show three distinct regions: a hydrophobic region at the amino terminus that functions as an ER signal, followed by a hydrophilic region, and then finally a short hydrophobic region just before the amino terminus of the putative mature protein (Figs 2, 4). This pattern is very similar to the pattern described for *psbO* (Ishida and Green 2002), and is consistent with function as transit peptides (Fig. 3). Different transit peptides for the same gene imply duplication within the nuclear genome after the acquisition of the transit peptide, or multiple chloroplast to nucleus transfer events. Another otherwise

exclusively plastid encoded gene, *rps2*, does not have an apparent ER signal sequence, even though a full-length sequence was obtained (Fig. 4).

Nucleus to Nucleus Gene Transfer

The dinoflagellate EST data suggest that in these organisms there has been massive transfer of chloroplast genes to the nucleus (Tables 1, 2). Although transfer of organellar genes to the nuclear genome is a well documented phenomenon, there are distinct patterns of gene content within lineages (Palmer and Delwiche 1998). In particular, all known plastids of red algae and secondary plastids derived from them have a relatively rich set of genes (Fig. 1), and from this it is possible to make inferences about the likely gene content of the ancestral dinoflagellate plastid. The distribution of endosymbiont genes among plastid and nuclear genomes cannot be known with certainty, but it is likely that many of the plastid-associated genes identified here had been transferred to the nuclear genome of the red algal symbiont prior to its acquisition by a dinoflagellate.

To place the scale of this transfer in perspective, analysis of the *Arabidopsis* nuclear genome found ~4500 genes that are likely to be of cyanobacterial (i.e., plastid) origin, accounting for roughly 17.6% of all protein-coding sequences (Martin et al. 2002). Chloroplast targeting sequences were found on well over 2000 genes (Arabidopsis Genome Initiative 2000). This corresponds fairly well to the known sizes of cyanobacterial genomes with 3168 genes in *Synechocystis* and 5368 genes in *Nostoc* (Kaneko et al. 1996, 2001), taking into account the fact that some of these genes have undergone duplication in the nuclear genome, and that not all genes of cyanobacterial origin are expressed in the plastid. It is clear that substantial reduction has occurred in all plastid genomes and has been an ongoing process (Palmer and Delwiche 1998). However, this reduction has a limit: when the known photosynthetic plastid genomes are compared, a set of 44 protein-coding genes are always plastid encoded (Fig. 1; Martin et al. 1998, 2002). In red algae and lineages with plastids derived from them, such as the cryptophytes and the heterokonts, chloroplast genomes are relatively large and complex, with a shared set of about 120 protein-coding genes (Douglas and Penny 1999). Thus, assuming that the peridinin-type plastid is indeed of red algal origin, it probably had a relatively rich starting set of genes and consequently a dramatic reduction in gene content.

Perhaps even more striking than the transfer of genes from the chloroplast to the nuclear genome – a well-documented process in the evolution of photosynthetic eukaryotes – is the presence within the EST data of many genes that are in the nuclear genome of both red algae and plants. These genes were probably transferred directly from the nuclear genome of the red algal chloroplast donor to the dinoflagellate recipient. While horizontal gene transfer among prokaryotes is now well documented, and transfer from prokaryotic genomes to those of eukaryotes is familiar in the context of organelles, transfer among eukaryotic nuclear genomes is not as well documented. Obligate cellular endosymbiosis is an extremely close relationship among organisms, and it is probably not surprising that gene transfer has been documented in several such cases. In cryptomonads there is evidence of large scale nucleus to nucleus gene transfer despite the presence of a vestigal red algal nucleus (Douglas et al. 2001), and it seems likely that similar transfer of genes will be found in organisms with secondary plastids that do not retain nucleomorphs. There is also evidence of at least one transferred gene in sea slugs that acquire and retain functioning plastids for a period of months (Pierce et al. 2003).

The LHC gene family seems to be a good example of nucleus to nucleus gene transfer from the dinoflagellate EST data. In all known organisms LHC genes are exclusively nuclear encoded. LHC sequences had previously been reported from *A. carterae*, and two of these were found to form a monophyletic group in phylogenetic analysis of LHCs from diverse algae, suggesting that the protein had diversified within dinoflagellates (Durnford et al. 1999). Our data revealed 11 members of this family that were previously unknown in dinoflagellates, indicating a broad diversity in the LHC family of dinoflagellates similar to the pattern found in plants (Durnford et al. 1999). Thus LHC diversity in dinoflagellates is more complex than had previously been appreciated.

Cyanobacterial Genes and Biochemistry

This survey found ESTs for several Calvin cycle genes, three of which were clearly recognizable as being cyanobacterial in origin: phosphoribulokinase, which is characteristic of the Calvin cycle, as well as transketolase and fructose-1,6-biphosphatase (Table 2), both of which function in the Calvin cycle, but are not exclusive to it. Another Calvin cycle protein, the carbon-fixing enzyme rubisco (ribulose-1,5-bisphosphate carboxylase/oxygenase), has had an unusual history of transfer in dinoflagellates, which are the only eukaryotes in which rubisco is encoded in the nuclear genome (as a single gene, *rbcL*), and it is an unusual form II (dimeric) rubisco

that is otherwise found only in anoxygenic proteobacteria (Morse et al. 1995; Rowan et al. 1996). While the origin of the dinoflagellate form II rubisco remains obscure, it is almost certainly not of cyanobacterial origin, and is an excellent example of horizontal gene transfer across domains (Delwiche and Palmer 1996). In addition to these Calvin cycle genes, genes encoding triosephosphate isomerase and fructose-1,6-biphosphate aldolase were also present and are necessary for the regeneration of ribulose, but these ESTs do not provide enough information to determine if these are cyanobacterial or cytosolic forms of the enzymes. A substitution of a cytosolic glyceraldehyde 3-phosphate dehydrogenase (GAPDH) in dinoflagellate chloroplasts has been documented (Fagan et al. 1998; Fast et al. 2001). It seems dinoflagellates are using a suite of cyanobacterial genes for some reactions of the Calvin cycle, but two key reactions, catalysed by rubisco and GAPDH rely on bacterial and cytosolic genes, respectively.

Many other genes of cyanobacterial (plastid) origin were found, including a nearly complete suite of chlorophyll biosynthesis genes. The carotenoid-biosynthesis genes identified were farnesyl pyrophosphate synthase from *L. polyedrum* and two different forms of violaxanthin de-epoxidase from *A. carterae*. Cyanobacteria and plastids synthesize heme from glutamate (Buchanan et al. 2000) and the *A. carterae* library had glutamate semialdehyde synthase in high abundance. While we cannot rule out a separate mitochondrial pathway in dinoflagellates, these data indicate that the cyanobacterial version of this pathway, involving glutamate is present and highly expressed.

Other plastid associated pathways include fatty acid biosynthesis and the DOXP/MEP pathway, and genes corresponding to both of these pathways were found. Four fatty acid biosynthesis genes were found in this project: *fabD*, *fabB, fadE2* and a probable ketothiolase. The DOXP/MEP pathway of ~~meval-onate~~ biosynthesis is also present because a homolog of the *gcpE* (*ispG*) gene was found in *A. carterae*.

There is a single EST with similarity to a “plastid mRNA binding protein” implicated in processing the 3′ ends of chloroplast mRNAs in cyanobacteria and plants. This EST could provide the starting point for elucidating the transcription and translation of minicircle-derived genes.

Comparison with Plasmodium

Dinoflagellates are thought to be the sister taxon to the Apicomplexa, and these groups along with the ciliates constitute the Alveolata. Two ESTs that have good blastx similarity between these dinoflagellates and *Plasmodium* may be alveolate specific proteins, since they have no other significant matches. Also, if the Leu/Phe-tRNA protein transferase is, as the blast search suggests, a bacterial enzyme that is present in alveolates (Gardner et al. 1998), then a gene transfer event before the radiation of the lineage is most likely.

Conclusions

The results of this relatively small-scale study have allowed us to make specific, testable hypotheses concerning the evolutionary history, molecular biology, and biochemistry of dinoflagellate plastids. It is also possible that the relatively rich plastid-associated gene content in the nuclear genome partially explains the diversity of plastids and photosymbiotic associations that occur in dinoflagellates. Although one might expect that components of the photosynthetic apparatus would be unlikely to function in an unrelated plastid, *in vitro* reconstitution of LHC complexes with allochthanous pigments has demonstrated energy transfer in such heterogeneous complexes (Grabowski et al. 2001). Another hypothesis is that the ability to transfer typically plastid-encoded genes to the nucleus documented here may allow dinoflagellates to rapidly transfer genes from novel endosymbionts.

Methods

Library Construction

The first library from *Lingulodinium polyedrum* (= G*onyaulax polyedra*), strain 70, was donated by David Morse of the University of Montreal (Chaput et al. 2002), and a second from *Amphidinium carterae* CCMP1314 was prepared in house.

The directionally cloned *L. polyedrum* library was amplified once in lambda hosts. The cDNA sequences were excised from the phage according to the manufacturer's (Stratagene, La Jolla, CA) directions and subsequently handled as plasmids in *E. coli*.

Amphidinium carterae CCMP1314 was cultured in Atlantic ocean seawater (~32 ppt), supplemented to become Guillard's F/2–Si medium (Andersen et al. 1997), at 20 °C with a 14 hr/10 hr L:D cycle at 24 µmol photons/$m^2 \cdot s$. Cultures were harvested in log phase growth (10^4–10^5 cells/ml) at four time points in the daily cycle: once 2 hours after the lights were turned on and three subsequent times at 6 hour intervals. Approximately 8 l of culture were harvested

by centrifugation, flash frozen in liquid nitrogen, and stored at –80 °C. For RNA isolation, the method of Chomczynski and Sacchi was used: 2 grams of cells were collected from each time point, and ground with a Polytron (Kinematica, Luzon) homogenizer in Tri Reagent (Sigma, St. Louis, MO) at a ratio of 2 grams of cells/25 ml reagent. The polyadenylated fraction was isolated using a poly-T cellulose column and the cDNA library was constructed according to the protocol described (Sambrook et al. 1989). Reverse transcription was performed with 1000u SuperScript II RNase H-RT (Invitrogen, Grand Isle, NY) and 40u RNAsin (Promega, Madison, WI), with 5 micrograms of polyA RNA and 50 pmol of *NotI* polyT primer, GACTAGTTCTAGATCGCGAGCG GCCGCCCT ×15 (Piao et al. 2001) incubated at 42.5 °C for one hour in a total volume of 100 microliters in a buffer of 50 mM Tris pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT. Second strand synthesis was performed at 15 °C with 75u T4 DNA polymerase, 25u *E.coli* DNA ligase, and 2u RNAase H (Invitrogen) for one hour in a 375 microliter volume in a buffer of 25 mM Tris HCl pH 7.5, 100 mM KCl, 5 mM MgCl, 10 mM $(NH_4)_2SO_4$, 0.15 mM β-NAD, 0.25 mM dNTPs. The cDNA was polished with *Pfu* polymerase (Stratagene) at 72 °C for 20 minutes in a 40 microliter volume, methylated with *EcoRI* methylase (New England Biolabs; NEB, Beverly, MA), ligated to a synthetic linker (NEB) with *EcoRI* sites, and double digested with *EcoRI* and *NotI*, followed by size fractionation through a sepharose CL-4B column (Amersham-BioSciences, Piscataway, NJ). The cDNA was then ligated to a modified pBluescript *EcoRI*, *NotI* gel isolated vector and transformed into XL-10 Gold competent cells (Stratagene). This library was not amplified in any way.

Sequencing: Plasmids from individual clones were isolated using the 'miniprep' procedure (Sambrook et al. 1989), and sequenced using dye terminator chemistry (ABI). For the *L. polyedrum* library the M13–20 primer was used for 5′, and T7 for 3′ sequencing. For the *A. carterae* library, a custom primer that ends at the *EcoRI* site of the linker was used for 5′ sequencing and M13–20 for 3′ sequencing. Reactions were performed at the reduced volume recommended for 384 well plates. The reactions were analyzed with an ABI 3100.

Bioinformatics: Sequences were edited using the program Sequencher (GeneCodes, Ann Arbor); vector and low quality bases were removed, and in some cases manual editing was used to restore low quality data, particularly when a poly-A tail was identified in the region of low quality sequence. Beginning and end of high quality data were also verified with phred (Ewing et al. 1998) to ensure consistency and promote automation. The individual ESTs were then exported to a FileMaker Pro (FileMaker, Santa Clara, CA) database and used individually for blast sequence similarity searches (Altschul et al. 1997).

Several searches were performed for each EST. Blastcl3 was used to perform blastn (nucleotide) and blastx (translated nucleotide) searches against the entire GenBank nr (nonredundant) database, as well as a blastx search limited to the entrez query *"Plasmodium,"* and a tblastx search against dbEST. Blastall was used to perform local blastn searches that reciprocally compared our two dinoflagellate EST databases. The results of these searches, as well as predicted translations were parsed using PERL scripts and exported to the database. Summary data are presented in Tables 1–3.

Sequencher (GeneCodes) was used to cluster related and redundant ESTs by taking advantage of its contig assembly function. This allowed identification of gene families and partially overlapping ESTs, the latter of which can be assembled into longer contiguous sequences. When overlapping EST reads were identified from a putative single transcript (using minimum overlap criteria of 40 bases and 70% identity), manual editing was performed to ensure that the assembled contig was reliable and maintained an open reading frame. Homologous sequences with less than 70% identity were presumed to be members of a gene family, and sequences with less than 40 bp overlap were not assembled even when they were identified by blast as candidates to have been derived from identical transcripts. A contig (or cluster) database was maintained in parallel with the EST database, and all contigs were subjected to the same blast searches as above.

For transit peptide prediction, amino acid alignments derived from blastx results were used to determine the approximate beginning of the mature protein, and Kyle-Doolittle hydropathy plots were constructed for the putative leader sequence. SignalP and chloroP were used to identify targeting peptides (Nielsen et al. 1997; Emanuelsson et al. 1999).

Acknowledgements

Supported in part by NSF grant MCB-9984284. We are grateful to David Morse for the *L. polyedrum* library, to E. Gantt for advice and participation in the project, to Frank Albert for developing hydropathy plot software and to M.V. Sanchez Puerta and J. Palmer for review of the manuscript, members of the Delwiche Lab for critical comments, and to the Alfred P. Sloan Foundation for providing seed resources.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary-DNA sequencing – expressed sequence tags and human genome project. Science **252**: 1651–1656

Altschul SF, Madden TL, Schaffer AA, Zhang JH, 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

Andersen RA, Morton SL, Sexton JP (1997) CCMP – Provasoli-Guillard National Center for Culture of Marine Phytoplankton. J Phycol **33**: Supplement

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **393**: 796–815

Barbrook AC, Howe CJ (2000) Minicircular plastid DNA in the dinoflagellate *Amphidinium operculatum*. Mol Gen Genet **263**: 152–158

Buchanan BB, Gruissem W, Jones RL (2000) Biochemistry and Molecular Biology of Plants. American Society of Plant Physiologists, Rockville, MD

Cavalier-Smith T (1993) Kingdom Protozoa and its 18 Phyla. Microbiol Rev **57**: 953–994

Chaput H, Wang Y, Morse D (2002) Polyadenylated transcripts containing random gene fragments are expressed in dinoflagellate mitochondria. Protist **153**: 111–122

Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem **162**: 156–159

Delwiche CF (1999) Tracing the thread of plastid diversity through the tapestry of life. Am Nat **154**: S164–S177

Delwiche CF, Palmer JD (1996) Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. Mol Biol Evol **13**: 873–882

Dodge JD, Greuet C (1987) Dinoflagellate Ultrastructure and Complex Organelles. In Taylor FJR (ed) The Biology of Dinoflagellates. Blackwell Scientific Publications, Oxford, pp 92–142

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu XN, Reith M, Cavalier-Smith T, Maier UG (2001) The highly reduced genome of an enslaved algal nucleus. Nature **410**: 1091–1096

Douglas SE, Penny SL (1999) The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. J Mol Evol **48**: 236–244

Durnford DG, Deane JA, Tan S, McFadden GI, Gantt E, Green BR (1999) A phylogenetic assessment of the eukaryotic light-harvesting antenna proteins, with implications for plastid evolution. J Mol Evol **48**: 59–68

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

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8**: 175–185

Fagan T, Hastings JW, Morse D (1998) The phylogeny of glyceraldehyde-3-phosphate dehydrogenase indicates lateral gene transfer from cryptomonads to dinoflagellates. J Mol Evol **47**: 633–639

Fast NM, Kissinger JC, Roos DS, Keeling PJ (2001) Nuclear-encoded, plastid targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Mol Biol Evol **18**: 418–426

Gajadhar AA, Marquardt WC, Hall R, Gunderson J, Ariztia-Carmona EV, Sogin ML (1991) Ribosomal RNA sequences of *Sarcocystis muris, Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. Mol Biochem Parasitol **45**: 147–154

Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C, Pederson J, Shen K, Jing JP, Aston C, Lai ZW, Schwartz DC, Pertea M, Salzberg S, Zhou LX, Sutton GG, Clayton R, White O, Smith HO, Fraser CM, Adams MD, Venter JC, Hoffman SL (1998) Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science **282**: 1126–1132

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature **419**: 498–511

Grabowski B, Cunningham FX, Gantt E (2001) Chlorophyll and carotenoid binding in a simple red algal light-harvesting complex crosses phylogenetic lines. Proc Natl Acad Sci USA **98**: 2911–2916

Graham LE, Wilcox LW (2000) Algae. Prentice-Hall, Upper Saddle River, NJ

Grzebyk D, Schofield O, Vetriani C, Falkowski PG (2003) The mesozoic radiation of eukaryotic algae: the portable plastid hypothesis. J Phycol **39**: 259–267

Hecht S, Eisenreich W, Adam P, Amslinger S, Klaus K, Bacher A, Arigoni D, Rohdich F (2001) Studies on the nonmevalonate pathway to terpenes: the role of the *GcpE (IspG)* protein. Proc Natl Acad Sci USA **98**: 14837–14842

Hiller RG (2001) 'Empty' minicircles and *petB/atpA* and *psbD/psbE (cytb559a)* genes in tandem in *Amphidinium carterae* plastid DNA. FEBS Lett **505**: 449–452

Hiller RG, Wrench PM, Sharples FP (1995) The light harvesting chlorophyll *a-c*-binding protein of dinoflagellates: a putative polyprotein. FEBS Lett **363**: 175–178

Ishida K, Green BR (2002) Second- and third-hand chloroplasts in dinoflagellates: phylogeny of oxygen-evolving enhancer 1 *(PsbO)* protein reveals replacement of a nuclear-encoded plastid gene by that of a haptophyte tertiary endosymbiont. Proc Natl Acad Sci USA **99**: 9294–9299

Ishida K, Cavalier-Smith T, Green BR (2000) Endomembrane structure and the chloroplast protein targeting pathway in *Heterosigma akashiwo* (Raphidophyceae, Chromista). J Phycol **36**: 1135–1144

Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res **8**: 205–213

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Kakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabita S (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res **3**: 109–136

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV (1998) Gene transfer to the nucleus and the evolution of chloroplasts. Nature **393**: 162–165

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA **99**: 12246–12251

Morse D, Salois P, Markovic P, Hastings JW (1995) A nuclear-encoded form II rubisco in dinoflagellates. Science **268**: 1622–1624

Nassoury N, Cappadocia M, Morse D (2003) Plastid ultrastructure defines the protein import pathway in dinoflagellates. J Cell Sci **116**: 2867–2874

Nielsen H, Engelbrecht J, Brunak S (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering **10**: 1–6

Nikaido I, Asamizu E, Nakajima M, Nakamura Y, Saga N, Tabata S (2000) Generation of 10,154 expressed sequence tags from a leafy gametophyte of a marine red alga, *Porphyra yezoensis*. DNA Res **7**: 223–227

Palmer JD, Delwiche CF (1998) The Origin and Evolution of Plastids and their Genomes. In Doyle JJ (ed) Molecular Systematics of Plants II. Kluwer Academic Publishers, Boston, pp 375–409

Peltier J-B, Friso G, Kalume DE, Roepstorff P, Nilsson F, Adamska I, van Wijk KJ (2000) Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. Plant Cell **12**: 319–341

Piao Y, Ko NT, Lim MK, Ko MSH (2001) Construction of long-transcript enriched cDNA libraries from submicrogram amounts of total RNAs by a universal PCR amplification method. Genome Res **11**: 1553–1558

Pierce SK, Massey SE, Hanten JJ, Curtis NE (2003) Horizontal transfer of functional nuclear genes between multicellular organisms. Biol Bull **204**: 237–240

Reith M, Munholland J (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol Biol Rep **13**: 333–335

Rizzo PJ (1987) Biochemistry of the Dinoflagellate Nucleus. In Taylor FJR (ed) The Biology of Dinoflagellates. Blackwell Scientific Publishing, Oxford, pp 143–173

Rizzo PJ, Noodén LD (1972) Chromosomal proteins in the dinoflagellate alga *Gyrodinium cohnii*. Science **176**: 796–797

Rowan R, Whitney SM, Fowler A, Yellowlees D (1996) Rubisco in marine symbiotic dinoflagellates: form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family. Plant Cell **8**: 539–553

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning a Laboratory Manual. Cold Spring Harbor Laboratory Press, Plainview, NY

Schein AI, Kissinger JC, Ungar LH (2001) Chloroplast transit peptide prediction: a peek inside the black box. Nucleic Acids Res **29**: art. no.-e82

van den Hoek C, Mann DG, Jahns HM (1995) Algae: an Introduction to Phycology. Cambridge University Press, Cambridge, UK

Wolters J (1991) The troublesome parasites – molecular and morphological evidence that Apicomplexa belong to the dinoflagellate-ciliate clade. Biosystems **25**: 75–83

Zhang ZD, Green BR, Cavalier-Smith T (1999) Single gene circles in dinoflagellate chloroplast genomes. Nature **400**: 155–159

Zuegge J, Ralph S, Schmucker M, McFadden GI, Schneider G (2001) Deciphering apicoplast targeting signals – feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. Gene **280**: 19–26