

## ABSTRACT

Title of Dissertation: THE ORIGIN OF THE DINOFLAGELLATE PLASTID

Tsvetan Radoslavov Bachvaroff, Doctor of Philosophy 2004

Dissertation directed by: Associate Professor Charles F. Delwiche

Department of Cell Biology and Molecular Genetics

The peridinin pigmented dinoflagellate chloroplasts are the result of a secondary endosymbiotic event between a photosynthetic eukaryote and a dinoflagellate. Dinoflagellate chloroplast and nuclear evolution were independent before this endosymbiotic event. To reconstruct the evolution of the dinoflagellate chloroplast, phylogenies were constructed with a chloroplast gene *psbB*. The gene phylogeny should reflect the evolution of the chloroplast and indicate the plastid donor lineage. Gene sequences derived from the dinoflagellate chloroplast were extremely divergent but suggested that the plastid donor could have been a haptophyte. In an attempt to find better genes for analysis and to further understand gene transfer about 4900 randomly selected expressed genes were sequenced from two dinoflagellates, *Lingulodinium polyedra* and *Amphidinium carterae*. From these genes, thirty typically plastid-encoded

genes were found, including eight otherwise known only from plastid genomes. Based on poly-A tails, gene families, and leader sequences these genes appear to be nuclear-encoded in dinoflagellates. This result suggests that dinoflagellate chloroplasts may have the smallest protein-coding potential yet known. These genes and the partially sequenced chloroplast genome of a haptophyte were used in a phylogenetic analysis. There is strong conflict between genes encoded in the chloroplast and those in the nucleus. The chloroplast genes suggest relationship between haptophyte and dinoflagellate plastids, while the nuclear-encoded genes suggest a relationship with heterokonts. Chromophyte plastid monophyly is supported by these data but the single origin of the chromophyte plastid from red algae does not mean that the host lineages are monophyletic. These results are consistent with at least two different scenarios: either dinoflagellates and haptophytes independently acquired a plastid from the heterokonts, or dinoflagellates acquired their plastids from haptophytes, who in turn acquired their plastids from heterokonts. The evolutionary rate of the remaining plastid-encoded genes was compared with formerly plastid-encoded genes. These relative rate tests revealed strong incongruence between minicircle genes, formerly plastid-encoded genes and genes that were likely to have been acquired from the nucleus of the plastid donor lineage.

THE ORIGIN OF THE DINOFLAGELLATE PLASTID

by

Tsvetan Radoslavov Bachvaroff

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2003

Advisory Committee:

Professor Charles F. Delwiche, Chair/Advisor  
Dr. Kirk Apt  
Professor Elisabeth Gantt  
Professor Charles Mitter  
Professor Eugene Small

© Copyright by  
Tsvetan Radoslavov Bachvaroff  
2004

## ACKNOWLEDGEMENTS

The author would like to acknowledge the following individuals for their contributions to this document. Torstein Tengs taught me valuable laboratory techniques and performed many sequencing reactions on my behalf. He also made substantial intellectual contributions to my understanding of plastid evolution. Dr. David Oldach of the Institute for Human Virology generously loaned equipment and provided encouragement. Professor Robert Rowan of the University of Guam challenged me to find better data. The *psbB* sequences from haptophytes were obtained from Kamran Shalchian-Tabrizi in Professor Jakobsen's lab at the University of Oslo, Norway. Kamran also contributed to my understanding of dinoflagellate nuclear phylogeny by generously sharing unpublished data. Three talented undergraduates participated in the EST data collection, Ernest Williams, Gregory Concepcion, and Carolyn Rogers. Greg also contributed substantially to the dinoflagellate EST database. A fourth undergraduate, Ali Tabatabai, worked on a dinoflagellate rubisco project not presented here. Finally, I would like to thank Maria Virginia Sanchez Puerta who has read and commented on many subunits of this document, and who has sequenced the organellar genomes of *Emiliana*.

## TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS.....</b>	<b>ii</b>
<b>LIST OF TABLES.....</b>	<b>v</b>
<b>LIST OF FIGURES.....</b>	<b>vi</b>
<b>CHAPTER I INTRODUCTION .....</b>	<b>1</b>
I.1. INTRODUCTION .....	1
<i>Features of dinoflagellates</i> .....	2
<i>The peridinin plastid</i> .....	2
<i>The other plastids of dinoflagellates</i> .....	4
I.2. SECONDARY ENDOSYMBIOSIS .....	7
<i>When did dinoflagellates acquire the peridinin plastid?</i> .....	8
<i>Features of primitive dinoflagellates</i> .....	9
<i>Eukaryotes related to dinoflagellates</i> .....	10
<i>How many plastid gains?</i> .....	11
I.3. THE OTHER CHROMOPHYTE GROUPS.....	13
<i>Chromophyte plastid vs. host phylogeny</i> .....	19
<b>CHAPTER II DINOFLAGELLATE EXPRESSED SEQUENCE TAG DATA INDICATE MASSIVE TRANSFER OF CHLOROPLAST GENES TO THE NUCLEUS.....</b>	<b>23</b>
II.1. ABSTRACT.....	23
II.2. INTRODUCTION .....	24
II.3. RESULTS.....	26
<i>Overview</i> .....	26
<i>Plastid-associated sequences</i> .....	28
<i>Signal peptides</i> .....	33
<i>Nucleus-to-nucleus gene transfer</i> .....	33
<i>Comparison to Plasmodium</i> .....	34
II.4. DISCUSSION.....	39
<i>Overview</i> .....	39
<i>Chloroplast to Nucleus gene transfer</i> .....	40
<i>Nucleus-to-nucleus gene transfer</i> .....	41
<i>Cyanobacterial Genes and Biochemistry</i> .....	44
<i>Comparison with Plasmodium</i> .....	45
<i>Conclusions</i> .....	46
II.5. MATERIALS AND METHODS.....	46
<i>Library Construction</i> .....	46
<i>Sequencing</i> .....	48
<i>Bioinformatics</i> .....	48
<b>CHAPTER III MULTI-GENE ANALYSES OF PLASTID RELATIONSHIPS WITHIN THE CHROMOPHYTE ALGAE .....</b>	<b>50</b>
III.1. ABSTRACT .....	50
III.2. INTRODUCTION.....	51
III.3. 2. RESULTS .....	53

2.1 Minicircle genes .....	53
III.2.2 Plastid genes from the dinoflagellate EST survey .....	60
2.3 Comparison of the minicircle dataset with the nuclear-encoded dataset .....	62
III.4. 3. DISCUSSION .....	71
3.1 Assessing monophyly of chromophyte plastids .....	71
3.1 Minicircle data .....	71
3.2 Genes that are nuclear-encoded in dinoflagellates .....	72
3.3 Implications and models of chloroplast evolution .....	74
III.5. 4. METHODS .....	76
4.1 Sequence acquisition .....	76
4.2 Alignments .....	77
<b>CHAPTER IV EVOLUTIONARY RATES OF DINOFLAGELLATE CHLOROPLAST GENES: HETEROGENEOUS RATES OF EVOLUTION BETWEEN PLASTID-ENCODED AND PLASTID-TARGETED GENES .....</b>	<b>79</b>
IV.1. ABSTRACT .....	79
IV.2. INTRODUCTION .....	80
IV.3. RESULTS .....	84
<i>Relative rate tests</i> .....	84
<i>Likelihood ratio tests</i> .....	85
IV.4. DISCUSSION .....	90
IV.5. MATERIALS AND METHODS .....	93
<b>CHAPTER V CHROMOPHYTE PLASTID PHYLOGENY: A PLASTID- ENCODED GENE SUGGESTS THAT THE PERIDININ DINOFLAGELLATE PLASTID ORIGINATED FROM HAPTOPHYTES .....</b>	<b>96</b>
V.1. ABSTRACT .....	96
V.2. 1. INTRODUCTION .....	97
V.3. 3. RESULTS .....	99
V.4. 4. DISCUSSION .....	115
4.1. Long branch effects and parametric bootstrapping .....	115
4.2. Peridinin dinoflagellate plastids and haptophyte plastids .....	119
4.3. Models of chloroplast origin .....	121
V.5. 2. MATERIALS AND METHODS .....	122
<i>Culture conditions</i> .....	122
2.2. Isolation of the <i>A. sanguinea psbB</i> .....	122
2.3. PCR Amplification .....	123
2.4. Phylogenetic analysis .....	123
<b>CHAPTER VI CONCLUSIONS .....</b>	<b>128</b>
VI.1. THE MYSTERY OF DINOFLAGELLATE CHLOROPLAST DNA .....	128
VI.2. PHYLOGENY: SOME TESTABLE HYPOTHESES .....	130
<b>REFERENCES .....</b>	<b>134</b>

## LIST OF TABLES

Table II.1 Dinoflagellate ESTs present in the <i>Porphyra</i> plastid genome, sorted by bitscore .....	29
Table II.2 Plastid-associated genes not in <i>Porphyra</i> plastid genome (putative nuclear-to-nuclear transfers), sorted by bitscore.....	30
Table III.1 Useable alignment size and bootstrap proportions for selected groups using genes that are encoded on minicircles in dinoflagellates .....	57
Table III.2 Bootstrap proportions for selected groups using genes that are encoded on minicircles in dinoflagellates .....	58
Table III.3 Bootstrap support and size of useable alignment using the set of plastid genes that are nuclear-encoded in dinoflagellates. ....	67
Table IV.1 Results of likelihood ratio tests.....	89
Table V.1 Results of the Approximately Unbiased test.....	107
Table V.2 Sources of algal tissue and DNA sequences .....	124
Table V.3 Primers used for amplification and sequencing of <i>psbB</i> .....	125

## LIST OF FIGURES

Figure I.1 Dinoflagellate features .....	3
Figure I.2 Dinoflagellate nuclear tree .....	6
Figure I.3 Heterokonts .....	16
Figure I.4 Cryptophytes .....	17
Figure I.5 Haptophytes.....	18
Figure II.1 Venn diagram showing plastid gene content .....	35
Figure II.2 Transit peptide hydropathy plot .....	36
Figure II.3 Transit peptide hydropathy plot .....	37
Figure II.4 Transit peptide hydropathy plot .....	38
Figure III.1 Distribution of chloroplast genes used in this study .....	54
Figure III.2 Phylogenetic trees using minicircle genes.....	55
Figure III.3 Phylogenetic trees using combined minicircle genes.....	59
Figure III.4 Genes that are nuclear-encoded in dinoflagellates that group dinoflagellates with the red lineage .....	63
Figure III.5 Genes that are nuclear-encoded in dinoflagellates that are ambiguous .....	64
Figure III.6 Genes that are nuclear-encoded in dinoflagellates that group dinoflagellates with cyanobacteria .....	65
Figure III.7 Genes that are nuclear-encoded in dinoflagellates that group dinoflagellates with green algae .....	66
Figure III.8 Analyses of concatenated genes that are nuclear-encoded in dinoflagellates	69
Figure IV.1 Gene flow in secondary endosymbiosis.....	83
Figure IV.2 Distance based relative rate tests .....	86
Figure IV.3 Relative rates using the Rhino program .....	88
Figure V.1 Maximum likelihood vs. P-distances for <i>psbB</i> .....	102
Figure V.2 Phylogenetic tree using <i>psbB</i> .....	103
Figure V.3 Trees used for the approximately unbiased test.....	108
Figure VI.1 A phylogenetic reconstruction of chloroplast evolution. ....	133

## **Chapter I Introduction**

### **I.1. Introduction**

Dinoflagellates are primarily unicellular plankton with a proclivity for endosymbiosis. They are important primary producers, especially in marine environments where they are found as free-living plankton and as the endosymbionts within corals and giant clams. However, dinoflagellates are mostly known to the public because aggregations of many dinoflagellates form masses referred to as red tides (Anderson 1994; Anderson et al. 1998). These red tides can emit toxins that kill fish, disrupt human recreation, and sometimes cause human health effects. Dinoflagellates, like some other planktonic protists, are both producers and consumers in the food web, sometimes at the same time (Gaines and Elbrächter 1987; Bockstahler and Coats 1993). Both photosynthetic and non-photosynthetic dinoflagellates are known to consume other protists. In fact, although most cultured dinoflagellates are photosynthetic, perhaps half of all dinoflagellates are not photosynthetic (Taylor 1987), and a few are parasitic on crabs, fish, copepods, and even other dinoflagellates (Chatton 1920; Cachon 1987; Coats 1999). Dinoflagellate diversity was revealed by an elegant technique where deep ocean waters were sampled, DNA was extracted from these samples, and the resulting sequences were used in phylogenetic analysis. Many of these uncultured deep ocean organisms appear to be dinoflagellates, or at least related to dinoflagellates (Lopez-Garcia et al. 2001; Moon-van der Staay et al. 2001).

### *Features of dinoflagellates*

A combination of several features distinguish most dinoflagellates from other eukaryotes (Fig.I.1). The nucleus is packed with permanently condensed chromosomes that appear to lack nucleosomes and histones (Spector 1984; Rizzo 1987). When the nucleus divides, the nuclear envelope does not break down and the spindle poles form outside of the nucleus (Kubai and Ris 1969; Ris and Kubai 1974). The individual spindles pierce the nuclear envelope and segregate the chromosomes. The DNA is also extensively modified with 5-hydroxymethyluracil substituted for thymine (Rae and Steele 1978). DNA containing 5-hydroxymethyluracil is otherwise known only from phages.

### *The peridinin plastid*

There are many different chloroplast types in dinoflagellates, but the majority of photosynthetic dinoflagellates are pigmented with peridinin and chlorophyll *c* (Dodge 1975). Peridinin plastids contain the carotenoid peridinin which is unknown from any other organism (Fig.I.1D). The structure of peridinin is highly unusual for a carotenoid, with 37 instead of the typical 40 carbon skeleton, and an unusual third epoxide ring (Strain et al. 1971). Apparently during the biosynthesis of peridinin three carbons are removed (Swift 1981). This pigment is used to harvest light in two distinct types of pigment protein complexes (Prezelin 1987). The first complex is homologous to other eukaryotic light-harvesting proteins and is bound in the photosynthetic membrane (Hiller et al. 1993). The second complex also acts to harvest light and transfer light energy to the reaction center, but is water-soluble. The structure, sequence and position of the protein in this water-soluble complex are not homologous to any other known proteins (Hofmann et al. 1996).



Another feature of the dinoflagellate peridinin plastid is the form II rubisco. First immunologic, then protein and DNA sequence studies indicated that the dinoflagellate carbon fixing enzyme, rubisco, was unlike other eukaryotic rubisco enzymes (Morse et al. 1995; Whitney et al. 1995; Rowan et al. 1996). Phylogenetic analysis of dinoflagellate rubisco suggests that it is a form II type otherwise unknown from oxygenic phototrophs (Delwiche and Palmer. 1996). An *in situ* immunologic study indicates that this form II rubisco is active in the peridinin type plastid (Jenks and Gibbs 2000). Thus, dinoflagellate plastids are unique in their pigment, light harvesting system and carbon fixing system.

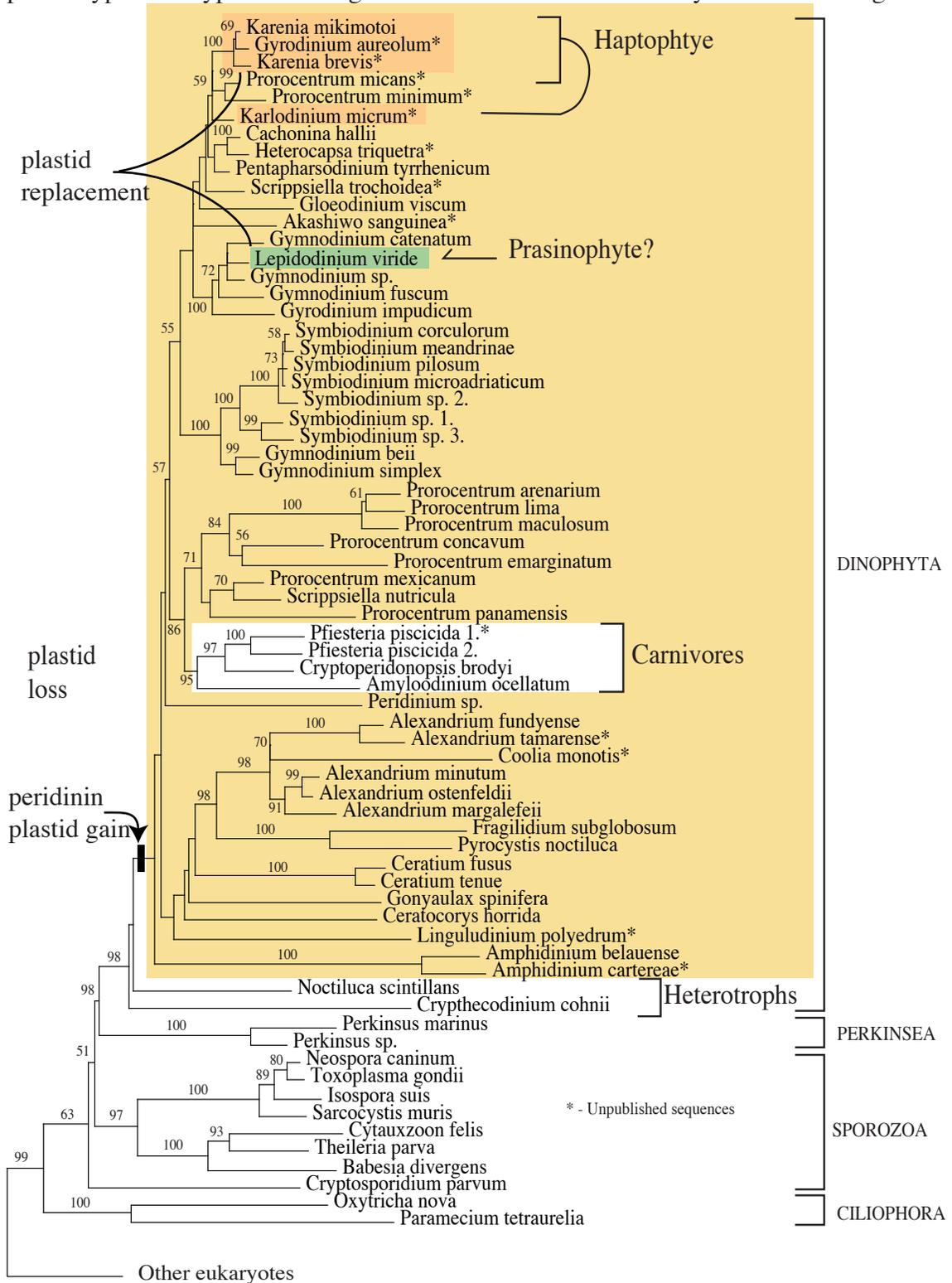
Other genes associated with the plastid are also unusual. Uchida (1988) showed that peridinin dinoflagellates have a divergent ferredoxin and later found that dinoflagellate *psbA* sequences are also extremely divergent from other algae (Takishita and Uchida 1999). Using density gradient centrifugation Zhang (1999) and later Barbrook and Howe (2000) as well as Hiller (2001), and Bachvaroff (unpublished) have separated circular, or plasmid-like DNAs from peridinin dinoflagellates. These minicircles encode one or sometimes two genes, although partial or incomplete genes have also been found (Hiller 2001). The messenger RNA for *psbA* is located in the plastid and therefore it is likely that the minicircles are also in the plastid (Takishita et al. 2003).

#### *The other plastids of dinoflagellates*

Dinoflagellates engage in a number of different endosymbiotic relationships, utilizing plastids from every major photosynthetic lineage. The ultrastructure of *Peridinium balticum* (= *Glenodinium foliaceum*) suggested that an entire eukaryote was inside the dinoflagellate host (Dodge 1971; Tomas and Cox 1973). Phylogenetic analysis

of plastid genes from the endosymbiont indicated that the endosymbiont is a diatom (Chesnick et al. 1996). There are also a group of dinoflagellates that appear to have engulfed a haptophyte. In these dinoflagellates, the nucleus of the endosymbiont is missing, but the plastid is pigmented with 19' hexanoyloxyfucoxanthin, a characteristic of haptophytes (Bjørnland and Tangen 1979; Tengs et al. 2000). Analyses of host and endosymbiont genes suggest that the 19' hexanoyloxyfucoxanthin dinoflagellates form a monophyletic group and contain plastids derived from haptophytes (Tengs et al. 2000). Other examples include dinoflagellates with green algal pigments, i.e. chlorophylls *a*, and *b* (Watanabe et al. 1990) or with cryptophyte pigments (Lucas 1990; Takishita et al. 2002; Hackett et al. 2003). In fact the number of plastid acquisitions within dinoflagellates may well exceed the number of endosymbiotic events in all other known eukaryotes, although it is not always clear if they are permanent endosymbionts (Takishita et al. 2002). The basis for frequent replacement of the peridinin type plastid is unknown (Saldarriaga et al. 2001). It may be due to deficiencies in the peridinin plastid, or to an increased ability to acquire endosymbionts in organisms that once had a peridinin plastid (Fig.I.2). In either case the abundance of secondary endosymbiotic events can be confusing.

Figure I.2 Dinoflagellate host phylogeny. Phylogenetic tree using the nuclear small subunit ribosomal RNA (18S). Fitch-Margoliash tree using LogDet distances. Putative plastid types for atypical dinoflagellates are shown. Tree courtesy of Torstein Tengs



## **I.2.Secondary endosymbiosis**

Secondary endosymbiosis is the transfer of an organelle, in this case the chloroplast, from one eukaryote to another. A transition from a predator - prey relationship to a stable symbiotic relationship is one way that a secondary endosymbiont could be established. In such a case, the digestion mechanism in the host must be altered so that the endosymbiont is not destroyed, and the host can obtain photosynthetic byproducts. In the ciliate *Myrionecta rubrum* the prey chloroplast and nucleus are retained for a period of weeks (Gustafson 2000). However, in this organism the survival of the ciliate seems to depend on the presence of the prey nucleus. There is a strong correlation between the decline in the host and the decline in the number of endosymbiont nuclei. For continued survival the *Myrionecta* must be fed fresh prey. To maintain the chloroplast without its nucleus some gene transfer is presumably required because at most only 200 chloroplast protein-coding genes are encoded in the plastids of known eukaryotes (Martin et al. 2002). The remaining genes, probably in the thousands, are nuclear-encoded (Martin et al. 2002).

Secondary endosymbiotic events also leave ultrastructural clues. For example, extra chloroplast membranes suggest that the plastid was either surrounded by a host-derived membrane, or retained an outer membrane of the endosymbiont (Gibbs 1962; Gibbs 1978; Gibbs 1981). In the case of the peridinin plastid a single extra membrane has been found around the plastid, but at present we do not know if this is a host or endosymbiont derived membrane. So, although most photosynthetic dinoflagellates contain this peridinin plastid, there are several other plastid types within dinoflagellates.

*When did dinoflagellates acquire the peridinin plastid?*

A good dinoflagellate host phylogeny is needed to determine when the peridinin plastid was acquired. This host phylogeny is used to infer when the plastid was gained, and when it was lost or replaced. For example, if all of the peridinin dinoflagellates formed a monophyletic group with the non-photosynthetic dinoflagellates as the outgroup then the acquisition of the peridinin plastid would be clear (Fig.I.2). Phylogenetic trees do show a non-photosynthetic outgroup to the peridinin dinoflagellates, but all of these trees suffer from low bootstrap support and consequently this result must be interpreted with caution (Litaker et al. 1999; Shalchian-Tabrizi 1999; Saldarriaga et al. 2001).

Saunders proposes an early gain of the dinoflagellate plastid in the blastodinium dinoflagellates (Saunders et al. 1997). The blastodinium dinoflagellates are parasites that reside inside the gut of copepods (Chatton 1920). They are pigmented and have been presumed to be photosynthetic (Saunders et al. 1997). However, the single attempt to measure photosynthetic activity in these dinoflagellates is not convincing since the amount of carbon fixed was minute (Pasternak et al. 1984). Another example is found in the curious dinoflagellates that prey on the filamentous green alga *Oedogonium* (Pfiester and Popofsky 1979). These dinoflagellates are remarkable for their ability to adopt an amoeboid form. Reports that these dinoflagellates are pigmented are probably accurate, but it seems likely that the pigment is derived from their prey, not from a stable endosymbiotic relationship (Cachon 1987). Since many of these parasitic or heterotrophic dinoflagellates are not cultured it is very difficult to absolutely reject the idea that they have plastids.

### *Features of primitive dinoflagellates*

The deepest branching dinoflagellates that have been cultured and characterized are *Amoebophyra*, *Oxyrris*, and *Noctiluca* and all are heterotrophic. *Amoebophyra* is parasitic on other dinoflagellates (Chatton 1920; Cachon 1987). Molecular trees place *Amoebophyra* at the base of the dinoflagellates, with or near the heterotrophic *Noctiluca* and *Oxyrris* (Gunderson et al. 1999; Litaker et al. 1999; Saldarriaga et al. 2003). Although *Amoebophyra*, *Noctiluca* and *Oxyrris* are very different from each other, some of their characteristics do appear to be primitive for dinoflagellates. *Noctiluca* has condensed chromosomes only during one phase of its life cycle (Pfiester and Anderson 1987) and *Amoebophyra* appears to lack condensed chromosomes as well (Cachon 1987). *Oxyrrhis* has enough basic nuclear protein to stain with the protein stain fast green (Cachon et al. 1979). These data suggest that the typical dinoflagellate chromosome is a shared derived character (synapomorphy) for all other dinoflagellates. The body shape of the *Amoebophyra*, *Oxyrrhis*, and *Noctiluca* swarmer stages is similar, with a poorly developed girdle (Chatton 1920; Cachon 1987; Pfiester and Anderson 1987). More derived dinoflagellates have a characteristic girdle containing the flagellum that encircles the cell. Thus the classical dinoflagellate body plan with a girdle and a second trailing flagellum may also be good synapomorphy for most dinoflagellates.

The systematics of dinoflagellates based on morphology has been based primarily on thecal plate arrangement. This armored layer of cellulose plates is a primary mechanism for identifying the different dinoflagellate species because it is visible under both light and electron microscopy. The Kofoid system of tabulation standardized the treatment of these plates (Kofoid 1909). Typically a Kofoid series begins at the anterior of the dinoflagellate and wraps in a spiral around the cell towards the posterior. This is

grossly similar to unwrapping a rhomboid sphere to create a two dimensional ribbon of polygons. The tabulation of dinoflagellates is the only accurate and widely used method of species identification (Steidinger and Tangen 1997). However, this system is difficult to apply to the ‘naked’ or athecate dinoflagellates, and often requires electron microscopy. Although naked dinoflagellates lack the cellulosic plates of the armored types, they do have a pattern of vesicles that are homologous to the thecae (i.e. thecal vesicles; Netzel and Dürr 1984). Moreover, some of these gymnodinoid (naked + dino) species now appear to be unrelated to each other based on molecular trees (Daugbjerg et al. 2000). This result suggests multiple independent losses of plates in dinoflagellates. Two major schools of thought have been applied to the evolution of dinoflagellate thecae. One concept involves a stepwise increase in plate number and complexity from the prorocentroid two-plate form onward to the complex many faceted shapes (Loeblich 1976; Taylor 1989). The other concept is that the prorocentroid form is derived by reduction (Eaton 1980). In fact Taylor (1987) proposes a rooted tree implying an increase in plate number that “would be somewhat the inverse of the present arrangement” (i.e. upside down) under the plate reduction hypothesis. While valuable insights have been gained from analyses of thecae it seems likely that the underlying premise of a stepwise increase or decrease in plate number is too simple to explain dinoflagellate morphologic diversity.

#### *Eukaryotes related to dinoflagellates*

The unique ultrastructure and division of the dinoflagellate nucleus led to a hypothesis separating dinoflagellates into a mesokaryote domain intermediate between eukaryotes and prokaryotes (Dodge 1965; Dodge 1966). While this mesokaryote

hypothesis has since been rejected it reveals the profound differences between dinoflagellates and other eukaryotes. It is all the more surprising then that the placement of dinoflagellates within other eukaryotes is well supported. Molecular analyses of ribosomal RNA and actin, as well as other genes, place dinoflagellates with ciliates (Lenaers et al. 1991) and the apicomplexans (Reece 1997; Saunders et al. 1997) . This group has been called the alveolates because most of these organisms share vesicles (alveoli) in their cell cortex. Although this synapomorphy is not absolute – some ciliates lack alveoli, and unrelated eukaryotes have similar structures – it remains a useful character. Two organisms have lately been added to the alveolates, *Perkinsus*, an oyster parasite, and *Parvilucifera*, a parasite on dinoflagellates, and are placed within the *Perkinsozoa* (Noren et al. 1999). The *Perkinsozoa*, apicomplexans, and dinoflagellates appear to be more closely related to each other than to the ciliates. In fact the boundary between basal dinoflagellates, apicomplexans and *Perkinsus* is unclear based on morphology, ultrastructure and molecular data. For example, the development of *Amoebophyra* spores involves sequential nuclear divisions within a single cytoplasm followed by the simultaneous partitioning of the cytoplasm between individual spores (Cachon 1987). This sporogenic division is almost exactly the same as the division of some apicomplexan parasites (Vivier and Desportes 1989).

#### *How many plastid gains?*

Major questions remain regarding the evolution of dinoflagellates, and the single gain of the peridinin plastid in typical dinoflagellates has been challenged by a recent discovery. Plastids were found in apicomplexans, even though apicomplexans are exclusively non-photosynthetic and parasitic (Egea and Lang-Unnasch 1995; Kohler et

al. 1997; Wilson 1997). These non-photosynthetic organelles are essential for cell division and play a role in some specific metabolic pathways (Fichera and Roos 1997; Gardner et al. 1998; Gardner et al. 2002). One possibility is that dinoflagellates and apicomplexans gained a plastid prior to the diversification of the host lineages, and that either plastids have been lost or are cryptic in the non-photosynthetic dinoflagellates (Zhang et al. 2000).

This hypothesis can be taken one step further, to include the heterokonts. There is some evidence from a multi-gene phylogeny that alveolates and heterokonts are related (Baldauf et al. 2000). The gain of the heterokont and alveolate plastid could then have occurred once, and plastids (or photosynthesis) were subsequently lost from the basal heterokont lineages, as well as from ciliates and basal dinoflagellates. This logic could be further extended to include organisms with plastids that are similar to dinoflagellate and heterokont plastids.

The modern version of this hypothesis has been called the chromalveolate hypothesis (Cavalier-Smith 1999; Cavalier-Smith 2000; Fast et al. 2001). In its first fully articulated form Christensen (1962) used plastid pigmentation to create a single group of chlorophyll *c* containing organisms, the chromophytes before secondary endosymbiosis was even recognized as a possibility. The four lineages of chlorophyll *c* containing plastids are: dinoflagellates, heterokonts, haptophytes and cryptophytes. In the intervening 40 years the number of predicted endosymbiotic events within these groups has varied between one (Cavalier-Smith et al. 1994) and four (Cavalier-Smith 1999). The primary factor has been the recognition of endosymbiosis as a major force in eukaryotic evolution (Sagan 1967); both mitochondria and primary plastids are now

understood to be endosymbiotic organelles (Margulis 1971; Gray and Spencer 1996). In primary plastid containing lineages, such as the green and red algae, the evolution of plastids and hosts are congruent (Delwiche et al. 1995). The same logic has been applied to mitochondria, where mitochondrial genes are often used to reflect host phylogeny (Burger et al. 1999; Lang et al. 1999). In primary lineages, there is no reason to separate the evolution of the different compartments. For example chloroplast genes are frequently used singly or in combination with mitochondrial and nuclear genes to resolve relationships within Plantae (Karol et al. 2001). But with the recognition that the chromophyte plastids are all enclosed by extra bounding membranes a separation of host and endosymbiont phylogeny became plausible (Gibbs 1978; Gibbs 1993; Palmer and Delwiche 1998). Clearly the evolution of a horizontally acquired organelle is not expected to be congruent with the evolution of the host, at least prior to the endosymbiotic event. This concept, along with the development of strong ultrastructural characters defining the chromophyte host lineages suggested multiple plastid gains by distinct host lineages (Delwiche 1999; Patterson 1999). Without strong host and chloroplast phylogenies the number of plastid gains will remain controversial.

### **I.3.The other chromophyte groups**

The diversity of the other three members of the chromophyte lineage is also substantial. The heterokonts (Fig. I.3) are a very large and diverse group that ranges from unicellular plankton and diatoms to the large brown algae (Patterson 1999; Patterson 1989). Within photosynthetic heterokonts there are several different pigment patterns, but the majority of heterokonts are pigmented with chlorophyll *c* and fucoxanthin (Van den Hoek et al. 1995). Photosynthetic heterokonts have plastids surrounded by an extra

host-derived membrane called the chloroplast ER because it is derived from the same layer of membrane as the ER (Gibbs 1981). The plastids have an extra band of three thylakoids surrounding the other thylakoid membranes called a girdle lamella (Van den Hoek et al. 1995). Non-photosynthetic and possibly aplastidic basal heterokonts include the labyrinthulomycetes and oomycetes (Van de Peer et al. 1996; Van de Peer and De Wachter 1997). Patterson (1989; 1999) also argues for an affinity of the slopalinids (*Opalina*) with the other heterokonts. The heterokonts are named for their two different flagella, one ornamented with tripartite hairs and the other naked. Most heterokonts also share a transitional helix at the base of the flagella. Oomycetes, labyrinthulomycetes and opaliniids could all be primitively aplastidic and plastids could have been gained by a heterotrophic member of the heterokonts, or these non-photosynthetic lineages could all have lost plastids in parallel and plastids were gained prior to the diversification of heterokonts.

A similar pattern is also true for the cryptophytes (Fig.I.4) in which the deepest branching lineage is heterotrophic and not photosynthetic (McFadden et al. 1994; Marin et al. 1998). Cryptophytes are remarkable because within the chloroplast ER is a relict red-algal nucleus, the nucleomorph (Gillot 1980). The complete nucleomorph sequence is known and it codes for a small number of plastid genes (Douglas 2001). Most plastid genes are encoded by the new host cryptophyte nucleus, and were probably transferred from the red algal nucleus. Cryptophyte pigmentation is in some ways intermediate between red algae and heterokonts. Biliproteins are present, like the red algae, but are not organized into phycobilisomes and are found within the thylakoids not outside of them (Gantt et al. 1971). Cryptophytes do have a light harvesting system that is similar

to heterokonts, but they utilize the xanthophyll alloxanthin, not fucoxanthin (Bjørnland and Liaaen-Jensen 1989).

The final group of chromophytes, haptophytes (Fig.I.5), is not very diverse in morphology, but is very important ecologically because of massive haptophyte blooms (Thomsen et al. 1994). The most common haptophytes have calcified scales called coccoliths; these have accumulated to form huge cretaceous chalk cliffs (Young et al. 1994). Open ocean blooms of haptophytes are large and affect global climate both by sequestering carbon as calcium carbonate and by emitting dimethyl sulfide, a chemical that can cause cloud nucleation (Westbroek et al. 1994). Haptophytes are pigmented with chlorophyll *a*, *c* and fucoxanthin, but also contain 19' linked fucoxanthin derivatives like 19' hexanoyloxyfucoxanthin (Bjørnland and Liaaen-Jensen 1989; Jeffrey and Wright 1994). The chloroplast is also surrounded by a chloroplast ER but lacks the girdle lamella of heterokonts. The defining feature of haptophyte host cells is a tentacle-like organelle called the haptonema (Parke et al. 1955; Van den Hoek et al. 1995). Haptophytes, unlike heterokonts, cryptophytes, and dinoflagellates, do not have a well documented non-photosynthetic basal lineage, although some haptophytes are heterotrophic (Marchant and Thomsen 1994).

The relationship of haptophytes to other eukaryotes is not certain. The haptophytes do not appear to be embedded within the heterokonts or sister to any other chlorophyll *c* containing algae based on mitochondrial gene trees (Sanchez Puerta et al). Moreover, position of the single haptophyte in this study, *Emiliania*, changed with different phylogenetic methods. Interestingly, the mitochondrial gene tree of host

Figure I.3: Heterokonts **A.** A schematic section through the heterokont *Ochromonas* (after Van den Hoek et al. 1995 figure 6.3). **B.** A nuclear small subunit ribosomal RNA tree is also shown with the potential plastid gain shown (after Van de Peers et al. 1997).

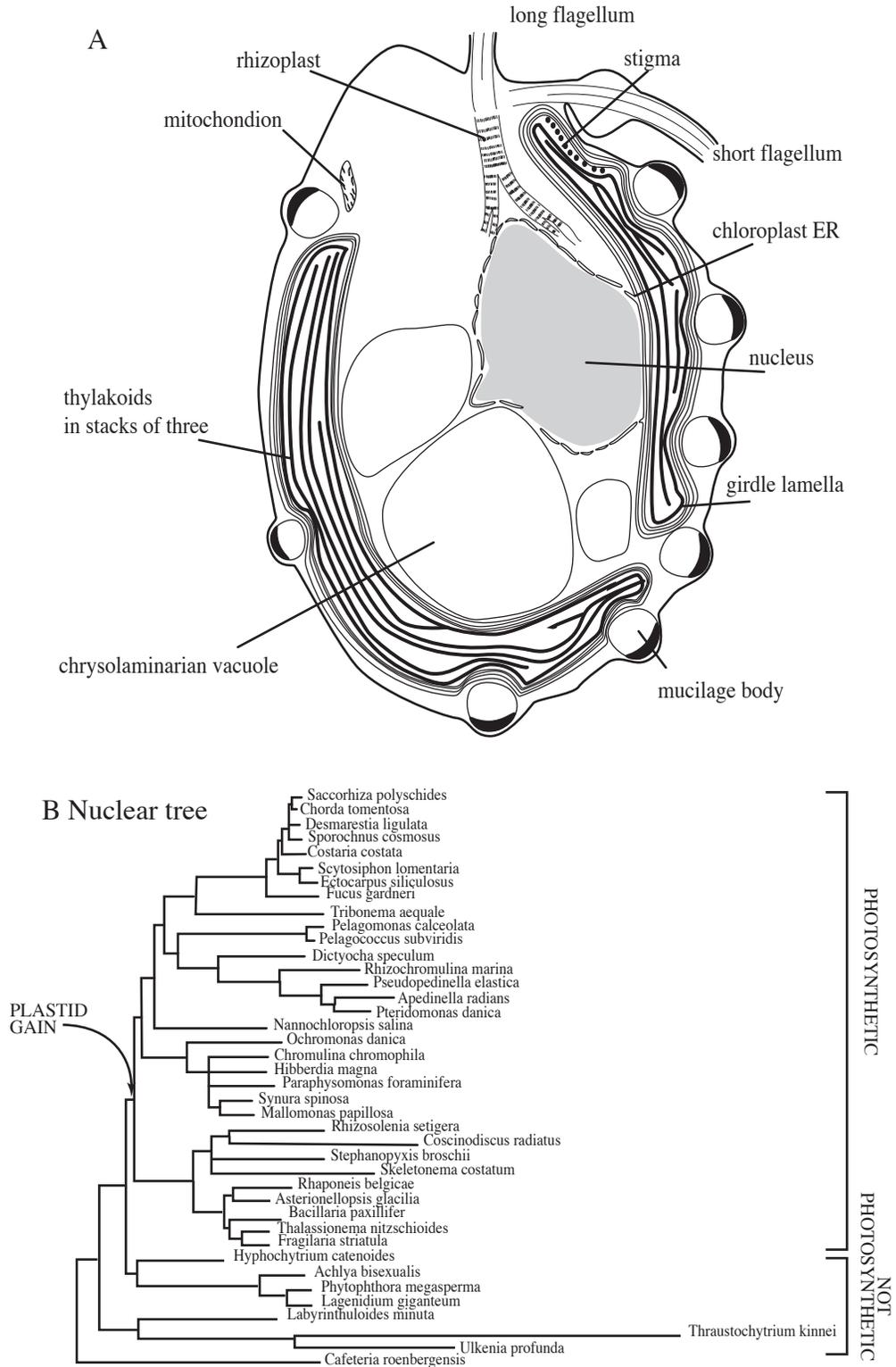


Figure I.3: Cryptophytes. **A.** A schematic section through a *Cryptomonas* cell shows the red algal endosymbiont enclosed within the chloroplast ER (after Van den Hoek 1995; figure 15.2). In cryptophytes a residual red algal nucleomorph is present. The thylakoid membranes are paired. **B.** A phylogenetic tree (after Marin and Melkonian 1998) is shown with the aplastidic *Goniomonas* as the deepest-branching cryptophyte. **C.** The homologous light-harvesting carotenoid, alloxanthin, is also shown.

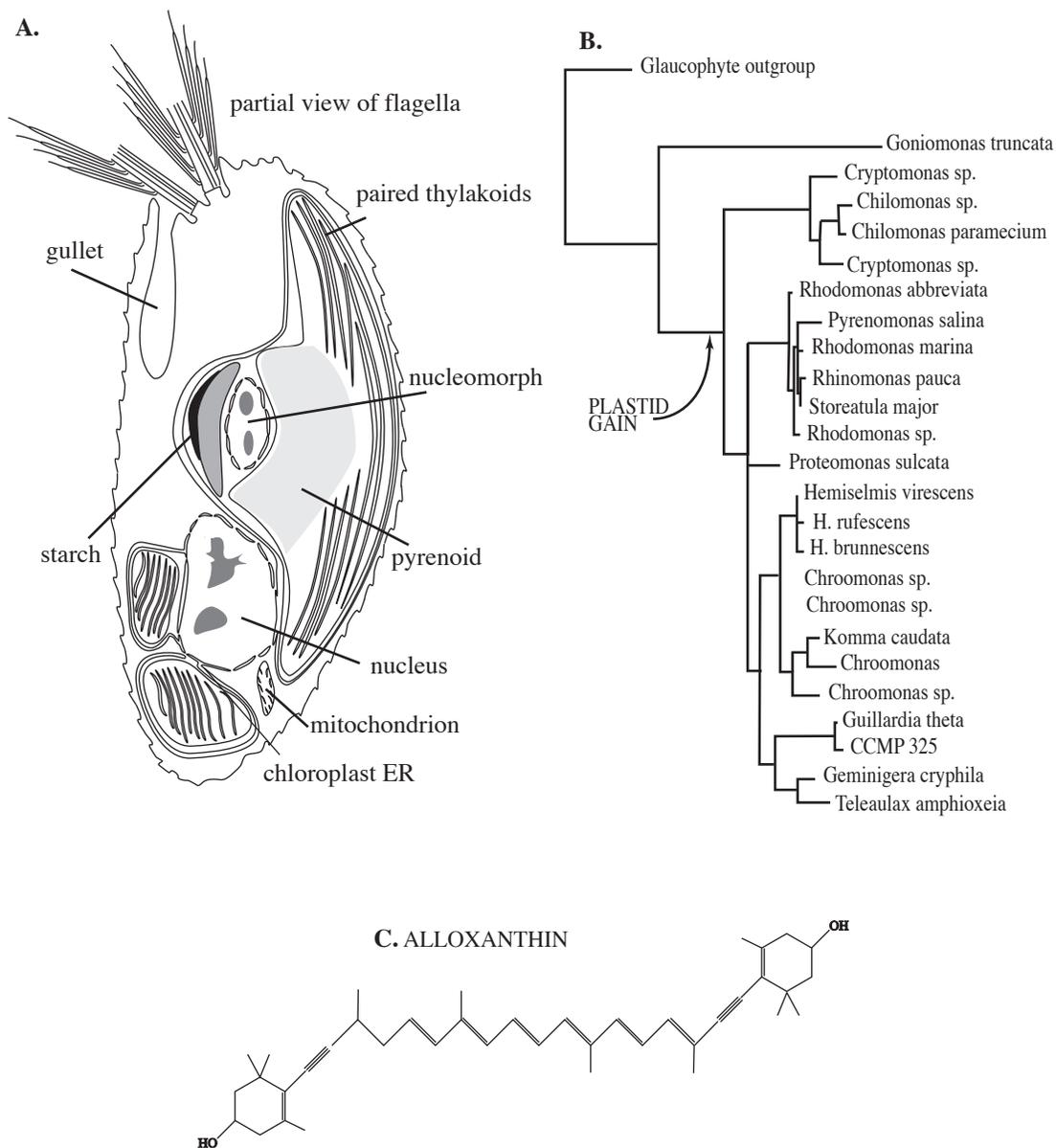
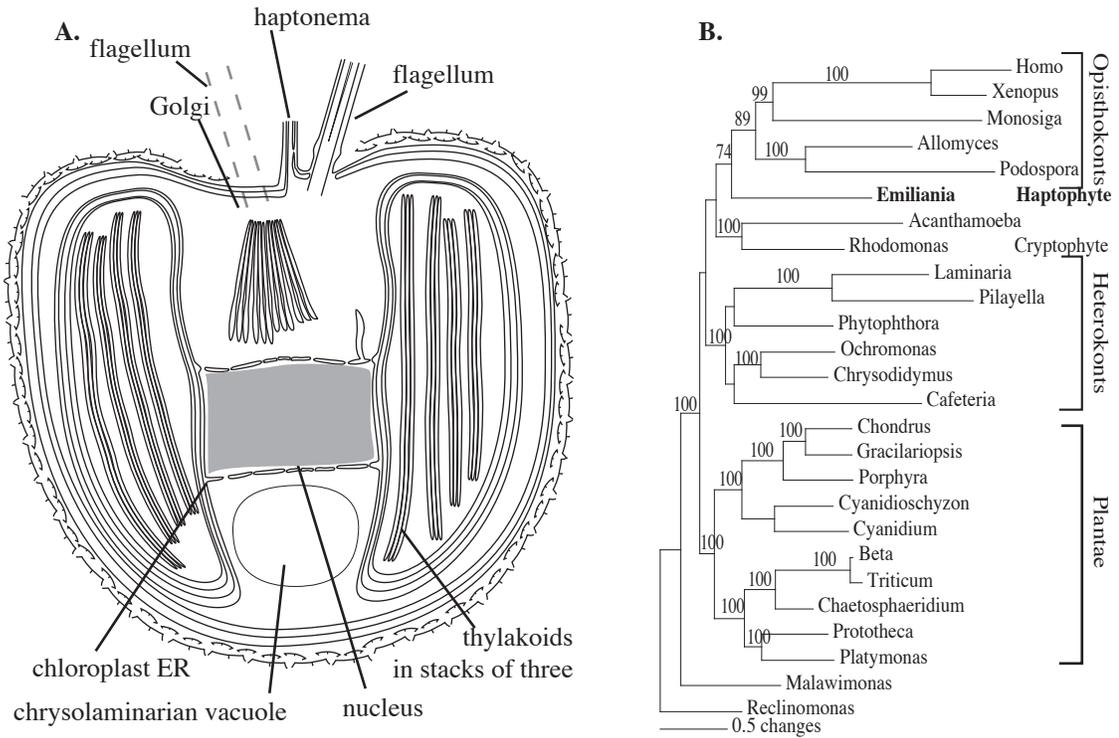


Figure I.5: Haptophytes. **A.** A schematic section through *Chrysochromulina* sp. is shown and the chloroplast ER, and thylakoids are shown (adapted from Van den Hoek et al. 1995; figure 14.2). **B.** A phylogenetic tree is also shown using mitochondrial genes from *Emiliana huxleyi* to infer host phylogeny (Tree courtesy of M.V. Sanchez-Puerta. Tree inferred by MrBayes using *cox1,2,3*, and *cob*. Posterior probabilities shown as percents above branches.) **C.** The homologous light-harvesting carotenoid, fucoxanthin, is shown.



phylogeny does recover a monophyletic *Plantae*, supporting a single origin of the primary plastid.

#### *Chromophyte plastid vs. host phylogeny*

Chromophyte plastid phylogeny is still controversial. The phylogeny of the plastids can always be made independent of host phylogeny by inferring additional independent endosymbioses. For example heterokonts, dinoflagellates and apicomplexans could have independently acquired plastids in three events. After the chromophyte plastid was gained once from a red alga it could have been passed to the other host lineages in which case the chromophyte plastid would appear to be monophyletic. Evidence for monophyly of the chlorophyll *c* plastid does not in itself indicate anything about the evolution of the host lineages, although this notion is accepted in the literature (Baldauf 2003; Palmer 2003). If plastids were gained once in the chromophytes then plastids were lost from all of the non-photosynthetic lineages within the proposed heterokont, alveolate, cryptophyte and haptophyte clade. Conversely, given the present data the phylogeny of the plastids and hosts can be made congruent by inferring additional losses of plastids. Eventually if the chromophyte plastid arose once and the hosts are monophyletic phylogenetic trees of host genes should become congruent with plastid trees.

The chromophyte algae do have several plastid features in common. First they all share chlorophyll *c*, a pigment not found elsewhere (Jeffrey 1989). Secondly chromophytes use xanthophylls like alloxanthin (cryptophytes) fucoxanthin (heterokonts and haptophytes) and peridinin (dinoflagellates) as light harvesting pigments (Bjørnland and Liaaen-Jensen 1989). The cryptophytes heterokonts and haptophytes also use a red

algal version of rubisco, while dinoflagellates use the odd form II rubisco (Delwiche and Palmer 1996). Phylogenetic analyses of light harvesting proteins also suggest a red algal origin of the chromophyte plastid (Durnford et al. 1999). All of these chromophyte lineages, as well as apicomplexans, use a cytosolic version of GAPDH in their plastids which argues against hypotheses that call for multiple plastid origins from the red algae (Fast et al. 2001; Harper and Keeling 2003).

While analyses of complete plastid genomes did not support the monophyly of heterokonts and cryptophytes (haptophyte plastid genome data were not available), both branched with the red algal lineage (Martin et al. 1998; Martin et al. 2002). Cryptophytes actually retain a highly reduced red algal nucleus within their periplastidal compartment, although few plastid targeted genes are still retained in that nucleus (Douglas 2001). A recent analysis of concatenated *psbA* and *psaA* genes suggested a haptophyte origin of the dinoflagellate plastid (Yoon et al. 2002). Moreover, the Yoon et al. (2002a) hypothesis specifies that the peridinin plastid originated from the lineage of dinoflagellates with haptophyte pigmentation. If this hypothesis is correct then the haptophyte containing dinoflagellates would be the outgroup to the peridinin dinoflagellates, but this placement is not seen in host trees (Litaker et al. 1999; Tengs et al. 2000; Saldarriaga et al. 2001). The chloroplast relationship using *psbA* and *psaA* is sensitive to the model of evolution used, and more sophisticated models imply separate gains of haptophyte plastids twice within dinoflagellates (Shalchian-Tabrizi 2003).

Another analysis of several plastid genes with good taxon sampling recovered a monophyletic chromophyte plastid lineage (Yoon et al. 2002). In this analysis dinoflagellates were not included, but cryptophytes formed a basal outgroup to the

heterokonts and haptophytes. While this analysis showed a single origin of the chromophyte plastid from red algae, this result does not eliminate the possibility of plastid transfer within the chromophyte host lineages. The phylogeny of the heterokont (Van de Peer and Wachter 1997) and cryptophyte (Marin et al. 1998) host lineages suggests that a single plastid gain could have occurred after the host lineages diversified. In both cryptophytes and heterokonts the most basal taxa are non-photosynthetic and may lack plastids. Moreover, a recent study of mitochondrial genes did not reveal a strong relationship of haptophytes to dinoflagellates or heterokonts (Sánchez Puerta et al. 2004), as would be predicted by both of these plastid gene analyses. A comprehensive understanding of both plastid and host phylogeny are needed to assess the evolution and number of endosymbiotic events within these algae.

The aim of this dissertation is to understand the evolution of the peridinin plastid within the context of the other chromophyte plastids. All of the dinoflagellates in the present study have peridinin and are unequivocally dinoflagellates. Dinoflagellates are difficult to work with, and the peridinin plastid particularly so; even now few data have been gleaned from this plastid. Some reasons for this are the extreme divergence of the minicircle genes, large-scale transfer of genes from the plastid to the nucleus and the potential for horizontal gene transfer. Because the peridinin plastid appears to contain few genuine plastid-encoded genes an expressed sequence tag approach was used both to understand gene transfer and content and to get useful genes for phylogeny. The analyses of these genes is challenging because of their extreme divergence, and in some cases transfer to the nucleus. Phylogenetic trees can be interpreted to be hypotheses about how evolution occurred but these hypotheses are only as strong as the model used to create

them; the assumptions of the model used to build the tree, the strength of the data, and the significance of the best tree with respect to other trees all need to be evaluated.

## **Chapter II Dinoflagellate expressed sequence tag data indicate massive transfer of chloroplast genes to the nucleus**

### **II.1. Abstract**

The peridinin 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 dinoflagellate plastid and its evolutionary history, an analysis was undertaken of arbitrarily selected sequences from cDNA libraries constructed from the peridinin-pigmented dinoflagellates *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 plastid has a minimal genome, and provide data that contradict the hypothesis that there is an unidentified canonical genome in the peridinin 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.

## II.2.Introduction

Dinoflagellates are environmentally and economically important flagellates that are common in both freshwater and marine environments. About half of all dinoflagellates are photosynthetic and most photosynthetic dinoflagellates are pigmented with peridinin and chlorophyll *c*. Like 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 (Van den Hoek et al. 1995). The majority of photosynthetic dinoflagellates rely on the peridinin plastid, but a number of other plastid types are found within the group, apparently the result of several independent symbiotic events (Watanabe et al. 1990; Chesnick et al. 1996; Delwiche 1999; Tengs et al. 2000). The typical, peridinin plastid is pigmented with 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 plastid are a chloroplast genome that is thought to consist entirely of single-gene minicircles (Zhang et al. 1999; Barbrook and Howe 2000; Hiller 2001), 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 (Van den Hoek et al. 1995; Graham

and Wilcox 2000). 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 and Noodén 1972; Rizzo 1987). 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 (Gajadhar et al. 1991; Wolters 1991; Cavalier-Smith 1993).

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 intriguing particularly in view of the complex interactions between organellar and nuclear genome.

To study the incorporation of the peridinin-type plastid into its present host, 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 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.

## II.3.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 exhausted.

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-1 $\alpha$  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, four 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 the two libraries produced a non-redundant set of 118 candidate plastid-associated sequences. Of these, 30 genes (Table II.1) – most of which were identified from the *A. carterae* library – are encoded in the plastid genome of *Porphyra* (Table II.2; Reith and Munholland 1995). The remainder is presumed to be nuclear-encoded in *Porphyra* and most other taxa (Table II.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 II.1. The ESTs that represented genes that are encoded in the plastid genome in *Porphyra* (Table II.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.

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

**bitscore**

gene <sup>a</sup>	bitscore <sup>b</sup>	e value <sup>c</sup>	clone reference <sup>d</sup>	variation <sup>e</sup>	polyA <sup>f</sup>	SignalP <sup>g</sup>	ChloroP <sup>h</sup>	Source <sup>i</sup>	Accession
<i>chlI</i>	358	1.0x10 <sup>-119</sup>	AcContig[0857]	family	yes	0.205	0.568*	A	CF067189
<i>atpI</i>	282	7.0x10 <sup>-76</sup>	AcContig[1157]	family	yes	0.817*	0.489	A	CF065976
<i>chlL</i>	234	1.0x10 <sup>-128</sup>	AcContig[0737]	1	yes	0.395	0.559*	A	CF064591
<i>ycf16</i>	187	3.0x10 <sup>-75</sup>	AcContig[1099]	1	yes	0.212	0.494	A	CF064637
<i>rps2</i>	176	1.0x10 <sup>-42</sup>	AcContig[0749]	1	yes	0.093	0.451	A	CF064824
<i>petK</i>	160	4.0x10 <sup>-37</sup>	AcContig[0964]	family	yes	0.932*	0.548*	A	CF066266
<i>petF</i>	152	3.0x10 <sup>-36</sup>	AcContig[1605]	family	yes	0.589*	0.486	Both	CF067664
<i>psaD</i>	148	1.0x10 <sup>-34</sup>	AcContig[0733]	1	yes	0.836*	0.571*	A	CF064527
<i>rplI</i>	139	9.0x10 <sup>-32</sup>	AcContig[0762]	1	yes	0.247	0.471	A	CF064976
<i>rpl16</i>	138	1.0x10 <sup>-31</sup>	Ac1119	-	yes	0.743*	0.532*	A	CF064566
<i>psaC</i>	127	3.0x10 <sup>-28</sup>	AcContig[1109]	family	yes	0.823*	0.518*	A	CF066614
<i>rpl13</i>	114	1.0x10 <sup>-24</sup>	AcContig[1636]	1	yes	0.763*	0.441	A	CF066354
<i>petJ</i>	110	1.0x10 <sup>-24</sup>	Ac5812	family	yes	0.355	0.487	A	CF067105
<i>secA</i>	108	6.0x10 <sup>-41</sup>	AcContig[1437]	1	yes	N.A.	N.A.	A	CF066408
<i>psaF</i>	103	1.0x10 <sup>-23</sup>	Ac977	-	yes	0.290	0.449	A	CF067650
<i>rpl3</i>	97	5.0x10 <sup>-23</sup>	AcContig[1546]	1	yes	0.736*	0.552*	A	CF067587
<i>psaE</i>	87	1.0x10 <sup>-16</sup>	Ac6843	-	yes	0.567*	0.481	A	CF067821
<i>ftsH</i>	85	7.0x10 <sup>-16</sup>	Ac1454r	-	no	N.A.	N.A.	Both	CF064829
<i>atpH</i>	84	9.0x10 <sup>-16</sup>	AcContig[0805]	family	yes	0.879*	0.516*	Both	CF067275
<i>tsf</i>	81	3.0x10 <sup>-32</sup>	AcContig[1710]	1	no	0.040	0.427	A	CF067081
<i>atpG</i>	77	2.0x10 <sup>-13</sup>	Ac1899	-	yes	0.580*	0.494	A	CF065024
<i>rpl4</i>	75	6.0x10 <sup>-19</sup>	AcContig[1547]	1	yes	0.255	0.482	A	CF066238
<i>clpC</i>	69	1.0x10 <sup>-22</sup>	AcContig[1539]	1	no	N.A.	N.A.	A	CF065755
<i>rps1</i>	69	1.0x10 <sup>-17</sup>	AcContig[1662]	1	yes	0.379	0.445	A	CF065490
<i>atpD</i>	64	1.0x10 <sup>-09</sup>	Lp587	-	yes	N.A.	N.A.	L	CD810773
<i>rpl33</i>	63	3.0x10 <sup>-09</sup>	Ac6830	-	yes	0.725*	0.555*	A	CF067798
<i>psaJ</i>	41	0.007	Ac1256	-	yes	0.226	0.430	A	CF064650
<i>psbY</i>	38	0.046	Ac6675	-	yes	0.319	0.509*	A	CF067444
<i>psbL</i>	35	0.17	Ac6375	-	yes	0.699*	0.541*	A	CF067332
<i>psbK</i>	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*.

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

Gene Name	length <sup>a</sup>	bitscore <sup>b</sup>	e value <sup>c</sup>	Clone reference <sup>d</sup>	Signal <sup>P</sup>	Chloro <sup>P</sup>	Accession <sup>e</sup>
Ribulose biphosphate carboxylase-oxygenase	1166	711	0.0	LpContig[0435]	0.060	0.429	CD810786
Glutamate semialdehyde synthase	1585	622	1.0x10 <sup>-177</sup>	AcContig[0832]	0.714*	0.515*	CF067231
Phosphoenol pyruvate synthase	1329	606	1.0x10 <sup>-172</sup>	LpContig[0475]	N.A.	N.A.	CD810119
Glyceraldehyde 3-phosphate dehydrogenase	1207	590	1.0x10 <sup>-167</sup>	LpContig[0425]	0.900*	0.565*	CD810603
Fructose-1,6-bisphosphate aldolase classII	1481	571	1.0x10 <sup>-161</sup>	AcContig[1111]	0.271	0.525*	CF067629
Peridinin chlorophyll protein (PCP)	1326	560	1.0x10 <sup>-158</sup>	LpContig[0334]	0.351	0.487	CD809573
Light-harvesting complex (LHC)	1212	541	1.0x10 <sup>-153</sup>	AcContig[0799]	0.744*	0.571*	CF066495
Histidine-tRNA ligase archeal	1438	406	1.0x10 <sup>-112</sup>	AcContig[1033]	0.051	0.450	CF064869
Oxygen evolving enhancer protein ( <i>psbO</i> )	1182	399	1.0x10 <sup>-110</sup>	AcContig[0758]	0.878*	0.513*	CF067369
Coproporphyrinogen III oxidase	1246	360	3.0x10 <sup>-98</sup>	AcContig[0734]	0.169	0.552*	CF064552
Porphobilinogen synthase	1480	344	2.0x10 <sup>-97</sup>	AcContig[1562]	0.640*	0.557*	CD806269
Transketolase 1 chloroplast	720	335	3.0x10 <sup>-91</sup>	Ac1168	N.A.		CF064604
Uroporphyrinogen decarboxylase ( <i>uroD</i> )	1455	313	4.0x10 <sup>-84</sup>	AcContig[0828]	0.269	0.558*	CF066269
Malonyl CoA:ACP transacyl carrier ( <i>fabD</i> )	1306	297	2.0x10 <sup>-79</sup>	AcContig[0959]	0.381	0.586*	CF067271
Aconitate hydratase	556	292	2.0x10 <sup>-78</sup>	Lp146b	N.A.		CD809560
Violaxanthin de-epoxidase precursor	1498	284	2.0x10 <sup>-75</sup>	AcContig[1564]	0.902*	0.482	CF066890
Ferredoxin NADP reductase	758	272	9.0x10 <sup>-72</sup>	AcContig[1305]	0.174	0.557*	CF067646
Mg protoporphyrin methyltransferase ( <i>chlM</i> )	1030	263	3.0x10 <sup>-69</sup>	AcContig[0790]	0.288	0.483	CF066233
Phosphoribulokinase	1281	256	1.0x10 <sup>-73</sup>	AcContig[0779]	0.101	0.438	CF065476
Triosephosphateisomerase	1077	253	2.0x10 <sup>-69</sup>	AcContig[1664]	0.070	0.443	CF066220
Phosphoserine aminotransferase	771	250	3.0x10 <sup>-67</sup>	Ac5574	0.157	0.438	CF066962
Inorganic pyrophosphatase	836	248	1.0x10 <sup>-64</sup>	Ac4379	0.237	0.525*	CF066545
Mg chelatase subunit ( <i>chlD</i> )	805	245	2.0x10 <sup>-63</sup>	AcContig[0766]	0.566*	0.522*	CF066220
Putative nucleotide-sugar dehydratase	728	229	7.0x10 <sup>-64</sup>	Lp1334	0.743*	0.460	CD809551
Flavoprotein cyanobacterial hits only	704	208	5.0x10 <sup>-53</sup>	Lp4457	N.A.		CD810473
Hydroxymethylbilan synthase	1272	207	1.0x10 <sup>-76</sup>	AcContig[1677]	0.263	0.480	CF065576
UDPglucose-starch glucosyltransferase	1305	204	3.0x10 <sup>-51</sup>	AcContig[1286]	0.241	0.599*	CF065409
Starch phosphorylase H	609	204	6.0x10 <sup>-52</sup>	Lp1271	N.A.		CD809492
Iron superoxide dismutase	579	201	7.0x10 <sup>-51</sup>	Lp2142	0.133	0.431	CD809812
Plastid mRNA binding protein	771	200	2.0x10 <sup>-50</sup>	Ac1315ra	N.A.		CF064730
Fructose 1,6-bisphosphatase	776	187	2.0x10 <sup>-46</sup>	Lp1187r	N.A.		CD810868
2-oxoglutarate/malate translocator	757	181	8.0x10 <sup>-45</sup>	Ac580r	N.A.		CF067093
ATP synthase gamma chain ( <i>atpC</i> )	1000	180	3.0x10 <sup>-44</sup>	AcContig[0867]	0.705*	0.536*	CF065394
3-oxoacyl-(acyl-carrier-protein) ( <i>fabB</i> )	604	169	3.0x10 <sup>-41</sup>	AcContig[1342]	N.A.		CF065186
<sup>b</sup> GI:16125539 <i>Caulobacter crescentus</i>	717	164	1.0x10 <sup>-39</sup>	LpContig[0273]	0.417	0.474	CD810585
Glutamine syntase III	845	162	5.0x10 <sup>-39</sup>	AcContig[1175]	0.133	0.455	CF065585
Thioredoxin reductase	739	160	2.0x10 <sup>-38</sup>	AcContig[1400]	0.249	0.470	CF064926
Protease ( <i>clpP1</i> )	685	159	3.0x10 <sup>-38</sup>	Ac3109	0.217	0.537*	CF065849
Glutamyl tRNA synthetase cytosolic?	430	159	1.0x10 <sup>-38</sup>	Ac6685	N.A.		CF067450
Putative nitrate transporter	651	154	1.0x10 <sup>-36</sup>	Lp43	N.A.		CD810421
Methyltransferase	1245	153	7.0x10 <sup>-36</sup>	AcContig[1313]	0.048	0.464	CF066859
<sup>b</sup> GI:22987108 <i>Burkholderia fungorum</i>	607	151	7.0x10 <sup>-36</sup>	Lp42	N.A.		CD810372
Alanine aminotransferase	771	139	6.0x10 <sup>-32</sup>	AcContig[1243]	N.A.		CF066050
Farnesylpyrophosphate synthase	694	139	3.0x10 <sup>-32</sup>	LpContig[0522]	0.062	0.433	CD810787
Phosphoglycolate phosphatase	739	137	1.0x10 <sup>-31</sup>	AcContig[1076]	0.408	0.498	CF066481
Aspartyl protease? Chloroplast nucleoid binding?	649	136	2.0x10 <sup>-31</sup>	Ac4923	N.A.		CF066716
Isocitrate lyase	659	134	1.0x10 <sup>-30</sup>	AcContig[1372]	N.A.		CF065691
Pyrophosphatase	499	128	3.0x10 <sup>-29</sup>	Lp374a	0.158	0.468	CD810255
Putative CP membrane-associated 30 kD protein	582	127	8.0x10 <sup>-29</sup>	Ac6374	0.573*	0.531*	CF067331
cGI:27382321 <i>Bradyrhizobium japonicum</i>	726	125	8.0x10 <sup>-28</sup>	Ac2672r	N.A.		CF065545
Glutathione peroxidase	692	125	7.0x10 <sup>-28</sup>	Ac3739	0.065	0.435	CF066302
Cytochrome B6-F complex iron-sulfur subunit ( <i>petC</i> )646	124	124	9.0x10 <sup>-28</sup>	AcContig[1301]	0.913*	0.549*	CF067078
<sup>b</sup> GI:23039345 <i>Trichodesmium erythraeum</i>	1403	119	2.0x10 <sup>-25</sup>	AcContig[0819]	0.518*	0.456	CF065847
Ketothiolase	544	113	2.0x10 <sup>-25</sup>	Ac6987	N.A.		CF067524
Monodehydroascorbate reductase	913	111	1.0x10 <sup>-23</sup>	Ac1976r	N.A.		CF065064
<sup>b</sup> GI:16330484 <i>Synechocystis</i> PCC 6803	699	110	1.0x10 <sup>-23</sup>	Ac3932	N.A.		CF066369
Photosystem II protein psbU	617	101	9.0x10 <sup>-21</sup>	Ac4283	0.857*	0.478	CF066491
Cobalamin synthase cGI:17229243	730	99	8.0x10 <sup>-20</sup>	Ac1263r	N.A.		CF064660

Chaperone ( <i>dnaJ hsp40</i> )	552	97	2.0x10 <sup>-20</sup>	Ac1889	N.A.		CF065020
<sup>b</sup> GI:17979159 Arabidopsis	829	97	3.0x10 <sup>-19</sup>	Ac5807	0.518*		CF067099
Elongation Factor G	488	95	4.0x10 <sup>-19</sup>	Ac2516	N.A.		CF065438
Phosphoglycerate mutase ( <i>gpmB</i> )	831	92	7.0x10 <sup>-18</sup>	Ac5805	N.A.		CF067097
<sup>b</sup> GI:15242446 Arabidopsis	493	91	9.0x10 <sup>-18</sup>	Lp4069	N.A.		CD810334
Chaperone ( <i>cpn60 groEL</i> )	516	90	1.0x10 <sup>-17</sup>	Ac6963	0.348	0.486	CF067852
Carbonic anhydrase	641	90	3.0x10 <sup>-17</sup>	Lp1702	N.A.		CD809667
PEP/phosphate translocator-like protein	460	86	2.0x10 <sup>-16</sup>	Ac5934	N.A.		CF067200
FKBP-type peptidyl-prolyl cis-trans isomerase	858	85	1.0x10 <sup>-15</sup>	AcContig[0742]	0.555*	0.547*	CF067531
Acyl-CoA dehydrogenase ( <i>fadE2</i> )	607	82	4.0x10 <sup>-15</sup>	Ac6379	N.A.		CF067335
Phenazine biosynthesis protein	593	80	1.0x10 <sup>-19</sup>	Ac3034	N.A.		CF065787
Pyridoxamine 5-phosphate oxidase	747	79	4.0x10 <sup>-17</sup>	Ac2510	0.085	0.436	CF065431
<sup>b</sup> GI:16329601 Synechocystis sp. PCC 6803	567	79	5.0x10 <sup>-14</sup>	AcContig[1192]	N.A.		CF066047
Some dnaJ similarity + ferredoxin	1167	77	7.0x10 <sup>-13</sup>	AcContig[0840]	0.326	0.531*	CF064949
Ferredoxin component	739	73	4.0x10 <sup>-12</sup>	AcContig[0864]	0.782*	0.552*	CF067564
Thioredoxin	796	71	2.0x10 <sup>-11</sup>	Ac1329r	0.764*	0.468	CF064745
Peroxisome/chloroplast ascorbate peroxidase	1665	69	2.0x10 <sup>-10</sup>	AcContig[0879]	0.149	0.456	CF067296
<sup>b</sup> GI:18405058 Arabidopsis	702	67	2.0x10 <sup>-10</sup>	AcContig[0923]	N.A.		CF066120
<sup>b</sup> GI:22326972 Arabidopsis	767	67	4.0x10 <sup>-10</sup>	LpContig[0295]	0.147	0.531*	CD809939
Putative methionyl-tRNA synthetase	666	65	7.0x10 <sup>-10</sup>	Lp1215	0.202	0.508*	CD809452
<sup>b</sup> GI:22971207 ABC transporter Chloroflexus	747	63	5.0x10 <sup>-09</sup>	Ac4912	0.420	0.496	CF066709
<sup>b</sup> GI:17230532 Nostoc sp. PCC 7120	451	62	1.0x10 <sup>-09</sup>	Ac1473	0.197	0.495	CF064845
Photosystem II 11kd protein	781	57	2.0x10 <sup>-07</sup>	AcContig[0903]	0.435	0.551*	CF065401
ABC-type transport protein	592	57	2.0x10 <sup>-07</sup>	Lp1325	N.A.		CD809542
WD domain	538	55	7.0x10 <sup>-07</sup>	Ac3924	N.A.		CF066361
RNA-binding protein ( <i>cp33</i> )	378	53	1.0x10 <sup>-06</sup>	Ac2021	N.A.		CF065099
<sup>b</sup> GI:13812240 Guillardia nucleomorph	785	53	6.0x10 <sup>-06</sup>	Ac4855	N.A.		CF066666
<sup>b</sup> GI:16329535 Synechocystis sp. PCC 6803	825	52	1.0x10 <sup>-05</sup>	Ac1749	0.339	0.436	CF064947
Chloroplast 28 kDa ribonucleoprotein	589	46	2.0x10 <sup>-04</sup>	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.

g – Indicates best hit to putative protein of unknown or unverified identity; the NCBI GI (“GenInfo”) number is provided as a uniform identifier

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” 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 II.2. A similar pattern of differences in transit peptides was found for the genes *psbO* (Fig. II.3) and *psaC* where greater variation was present in the leader than in the mature protein (data not shown).

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.

### *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. 2000; Peltier et al. 2000; Schein et al. 2001; Zuegge et al. 2001; Nassoury et al. 2003). Signal peptides were detected in a greater proportion of proteins destined for the thylakoid membrane (8 out of 12 in Table II.1), than in non-thylakoid proteins (5 out of 15 in Table II.1), but exceptions occurred even when apparently full-length sequences were found (i.e. *psaF* in figure II.4). None of the targeting-prediction software tested consistently recognized all these leader sequences as targeting peptides (Table II.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 LHC diversity 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 =  $6.0 \times 10^{-57}$ ) 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 =  $3.0 \times 10^{-26}$ ) 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), 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).

Figure II.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 mimircircles. The dotted line indicates those genes that are found in these genomes as well as in *Cyanophora paradoxa*, *Nephrolepis olivacea*, *Euglena gracilis*, *Zea mays*, *Oryza sativa*, *Nicotiana tabacum*, *Spinacia oleracea*, *Oenothera alata*, *Pinus thumbergii*, *Marchantia polymorpha*, *Mesostigma viride*, and *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 mimircircle-encoded gene found in the EST project. Additional comparisons are presented by Grzebyk (2003) and Martin et al. (2002). Gene nomenclature follows Martin et al. (2002).

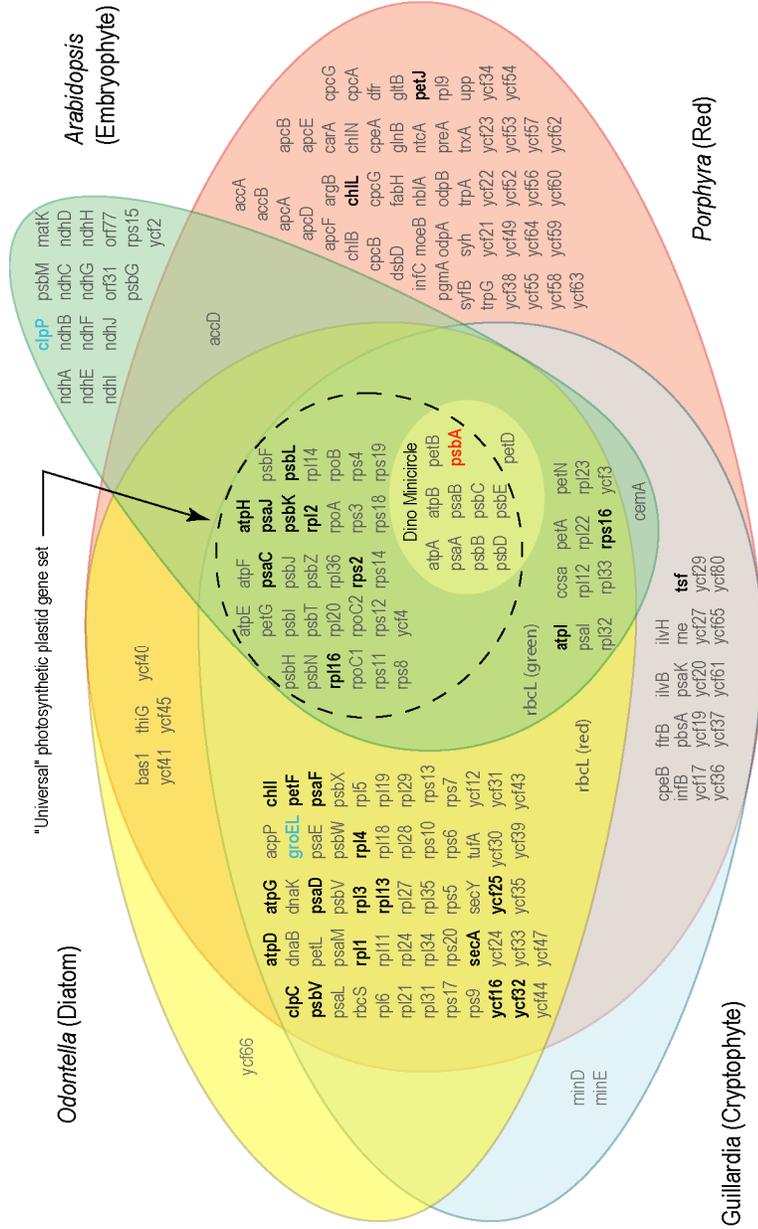


Figure II.2: Putative transit peptides from gene products of different *atpH* loci. Kyle-Doolittle hydropathy plots are shown with a window size of 7 amino acids. The chloroplast cleavage sites were inferred from an alignment of mature proteins. The predicted signal sequence cleavage sites are indicated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

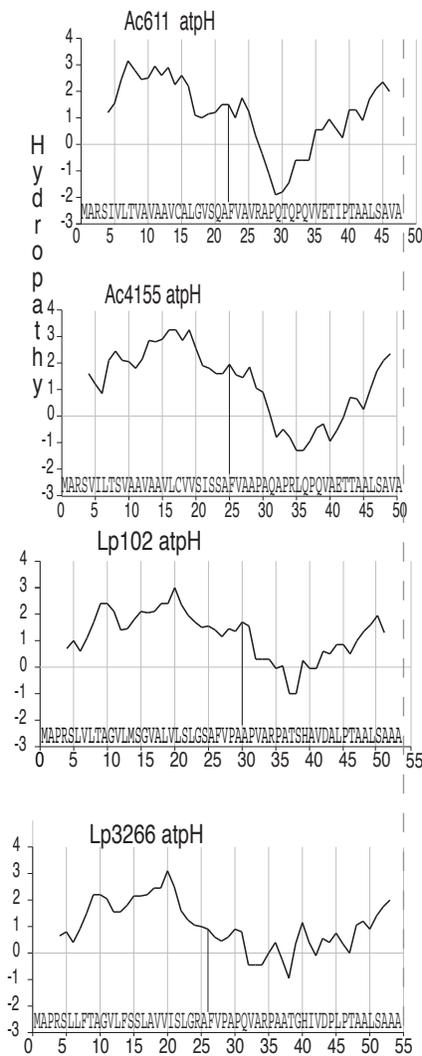


Figure II.3: Putative transit peptides from different copies of the *psbO* gene product. Kyle-Doolittle hydropathy plots are shown with a window size of 7 amino acids. The cleavage sites were inferred from a multiple sequence alignment. The chloroplast cleavage sites were inferred from an alignment of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

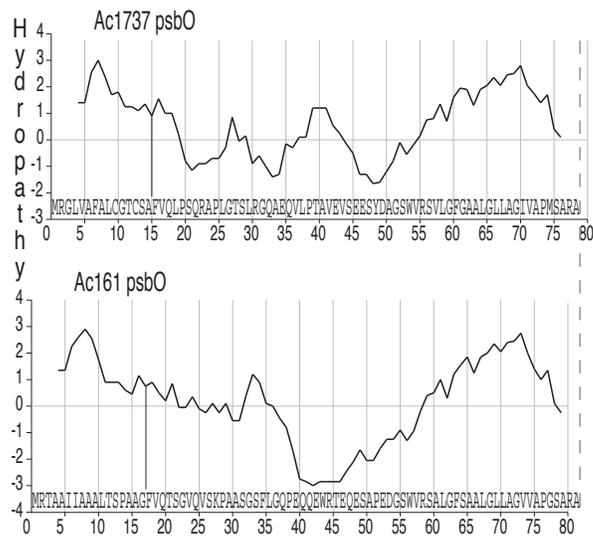
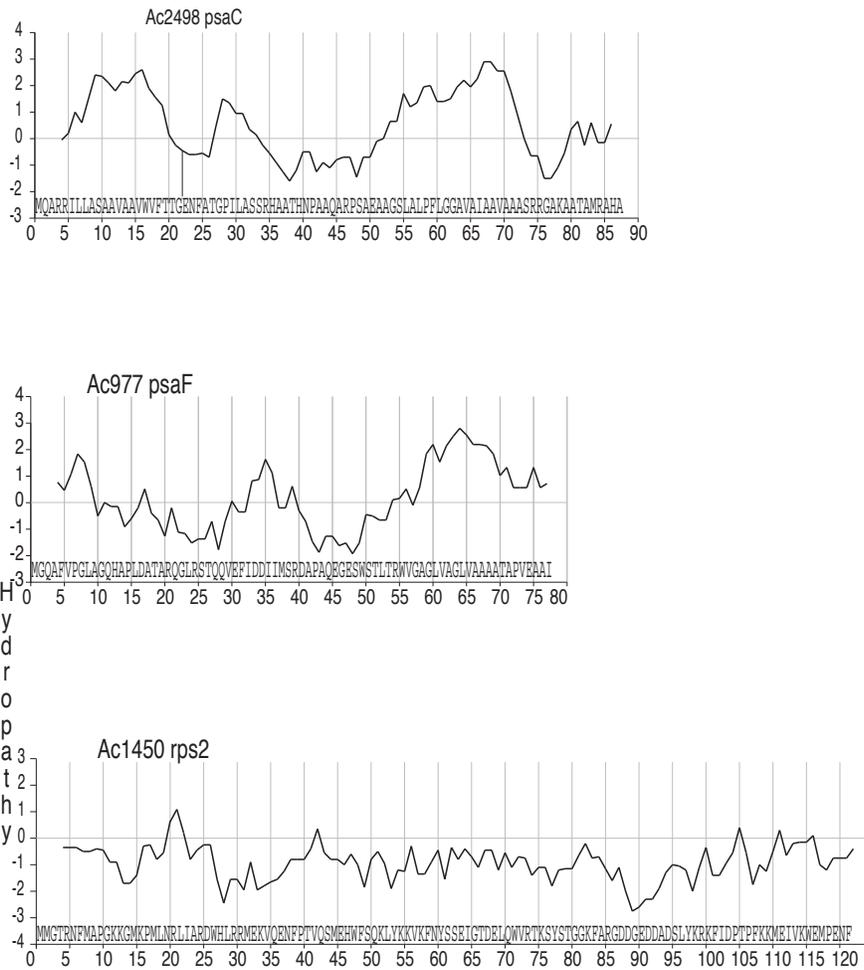


Figure II.4: Putative transit peptides from the *psaC*, *psaF*, and *rps2* gene products. Kyle-Doolittle hydrophathy plots are shown with a window size of 7 amino acids. The cleavage sites for these proteins were inferred from a multiple sequence alignment, although in the cases of the *psaF* and *rps2*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



## II.4.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 (Arabidopsis Genome Initiative 2000; 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 from *Amphidinium* 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 II.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 II.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 eukaryote 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 II.1 are homologous to plastid-encoded genes in *Porphyra*, and are almost certainly originally of plastid origin. Those in Table II.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. II.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. II.1, Table II.1), eight are present in all known photosynthetic chloroplast genomes (Gardner et al. 2002), and encode ribosomal proteins, ATP synthase, and photosystem components (Table II.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. II.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. II.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. II.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 II.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. II.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; Kaneko et al. 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 (Figure 1 Martin et al. 1998; Martin et al. 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 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 vestigial 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 II.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 isoprenoid 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 allochthonous 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.

## **II.5. Materials and Methods**

### *Library Construction*

The first library from *Lingulodinium polyedrum* (= *Gonyaulax 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 14hr/10hr L:D cycle at 24  $\mu\text{mol photons/m}^2\cdot\text{s}$ . 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 (1987) 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 / 25mL 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, GACTAGTTCTAGATCGCGAGCGGCCGCCCT x15 (Piao et al. 2001) incubated at 42.5°C for one hour in a total volume of 100 microliters in a buffer of (50mM Tris pH 8.3, 75mM KCl, 3mM MgCl<sub>2</sub>, 10mM DTT; Piao et al. 2001). 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 25mM Tris HCl pH 7.5, 100mM KCl, 5mM MgCl, 10mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 0.15mM  $\beta$ -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.

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).

## **Chapter III Multi-gene analyses of plastid relationships within the chromophyte algae**

### **III.1. Abstract**

The chromophyte algae are composed of four distinct lineages the dinoflagellates, haptophytes, heterokonts, and cryptophytes, all containing secondary plastids pigmented with chlorophyll *c* ultimately derived from the red algae. To better understand the plastid relationships of these algae multi-gene analyses of plastid-encoded or plastid-derived genes were performed. Overall 28 genes were used in this analysis with representatives of all four chromophyte lineages. Ten of these genes are encoded on minicircles in dinoflagellates. The other 18 are plastid-encoded in the red plastid lineage and nuclear-encoded in dinoflagellates. Analysis of the extremely long branch minicircle encoded genes suggests a relationship of haptophyte and dinoflagellate plastids and chromist monophyly. The *psbA* gene appears to be incongruent with the other nine genes in the minicircle dataset, and excluding *psbA* increases bootstrap support for chromophyte monophyly. However, analysis of 18 nuclear-encoded plastid derived genes from dinoflagellates produces conflicting results. A few of these genes support a relationship of dinoflagellates with green algae or cyanobacteria and are not congruent with the other gene trees. By excluding genes that do not support a red algal ancestry of dinoflagellates a compatible dataset was assembled. Using a subset of these genes a dinoflagellate relationship to heterokonts is shown with good bootstrap support, but the chromophytes are not monophyletic. While chromophyte plastids appear to have arisen once from the red algae, this result does not require the host lineages to be monophyletic since the plastid could have been transferred from one chromophyte host lineage to another.

### **III.2.Introduction**

The chromophyte, or chlorophyll *c* containing algae are an abundant and diverse group of organisms. Taken together this group of algae includes important coastal and oceanic primary producers, as well as harmful algae. The plastid relationships between the dinoflagellates, heterokonts, haptophytes and cryptophytes remain ambiguous. Recent evidence suggests that chromophyte plastids may be monophyletic, that is derived once from red algae (Fast et al. 2001; Yoon et al. 2002). However, the position of the dinoflagellate plastid in particular is unclear using typically plastid-encoded genes (Yoon et al. 2002). Dinoflagellates appear to have an extremely reduced plastid coding potential, and about eleven protein-coding genes that would normally be found in the chloroplast genome are found on single or double gene minicircles (Zhang et al. 1999; Barbrook and Howe 2000; Hiller 2001). Apparently many of the remaining genes have been transferred to the dinoflagellate nucleus (Bachvaroff et al. 2004).

Phylogenetic analysis of minicircle genes is difficult because of the extreme branch lengths leading to dinoflagellates (Takishita and Uchida 1999; Zhang et al. 1999; Zhang 2000; Yoon et al. 2002a; Chapter V). However, results from minicircle genes suggest a haptophyte origin, or at the least a sister relationship of haptophyte and dinoflagellate plastids (Yoon et al. 2002a; Shalchian-Tabrizi 2003; Chapter V). Analyses of light-harvesting complex genes suggested a relationship between heterokont and dinoflagellate plastids, and the branch lengths on this tree are more reasonable (Durnford et al. 1999). However, only a single gene was present from haptophytes and light-harvesting complexes are usually present in multiple gene families, so it is difficult to assess this result since non- orthologs could have been sampled. Analyses of

Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH), while providing strong support for plastid monophyly, produce jumbled trees within the chromophytes (Fast et al. 2001).

Obviously more data are needed to understand the exact relationships within the chromophyte plastids. The appropriate genes are now available from dinoflagellates, albeit in the form of nuclear-encoded plastid-targeted genes (Chapter II). To test the relationship of haptophyte, dinoflagellate and heterokont plastids, more data was needed from haptophytes. As part of an ongoing project to sequence the organellar genomes of *Emiliana huxleyi* (Sánchez Puerta et al. 2004), approximately two thirds of the chloroplast genome of this organism has been sequenced and phylogenetic analyses were performed using a portion of these data.

The taxon sampling (Fig. III.1) includes representatives of all four chromophyte lineages since the chloroplast genomes of a heterokont (Kowallik et al. 1995), and a cryptophyte (Douglas and Penny 1999) are already known. Several red algal plastid genomes are known including *Porphyra* (Reith and Munholland 1995), *Cyanidium* and *Cyanidioschyzon* (Ohta et al. 2003). The red algae and the chromophytes will be described here as the red algal plastid lineage. Chloroplast genomes are also available from diverse members of the green algae including the prasinophyte *Nephroselmis* (Turmel et al. 1999), and the streptophytes *Mesostigma* (Lemieux et al. 2000), *Chaetosphaeridium* (Turmel et al. 2002), and *Arabidopsis* (Sato et al. 1999). *Arabidopsis* is used here since several of the genes that are plastid-encoded in the red algal lineage are nuclear-encoded in the green lineage allowing for more comprehensive analysis. A glaucophyte plastid genome from *Cyanophora paradoxa*, is also known (Stirewalt et al.

1995), and complete cyanobacterial genomes are known from *Nostoc* and *Synechocystis* (Kaneko et al. 1996; Kaneko et al. 2001).

### III.3.2. Results

The distribution of the genes between chloroplast and nucleus are shown in Figure III.1. Phylogenetic trees for the 28 genes analyzed are shown in figures III.2-8 using the same scale within figures III.2, 3 and within figures III.4-8. Bootstrap values are shown when greater than 65%.

#### 2.1 Minicircle genes

These ten genes are all plastid-encoded in this dataset, and are characterized by extreme branch lengths leading to the dinoflagellate. Large portions of the *Amphidinium* sequences are difficult to align and are completely different from all other sequences. Results of phylogenetic analyses when dinoflagellates were excluded (trees not shown) show that only *atpB*, *petB* and *psbA* do not recover a monophyletic red lineage, here defined as red algae and secondary plastids derived from red algae (Table III.1). When dinoflagellates are included the bootstrap support for red lineage monophyly drops compared to the single gene analyses without dinoflagellates (Table III.2; Fig. III.2). Remarkably the gene that shows the strongest bootstrap support for the dinoflagellate + haptophyte clade, *psbA*, is also strongly incongruent with the other data, placing dinoflagellates and haptophytes with the outgroup cyanobacteria with strong bootstrap support. This result is supported by the Approximately Unbiased (AU) test (Shimodaira 2000) that strongly rejects dinoflagellate + red monophyly for the *psbA* gene. The AU test also rejects monophyly of *Amphidinium* + cyanobacteria because *Emiliania* is then excluded from that clade.

Figure III.1: Distribution of chloroplast genes used in this study

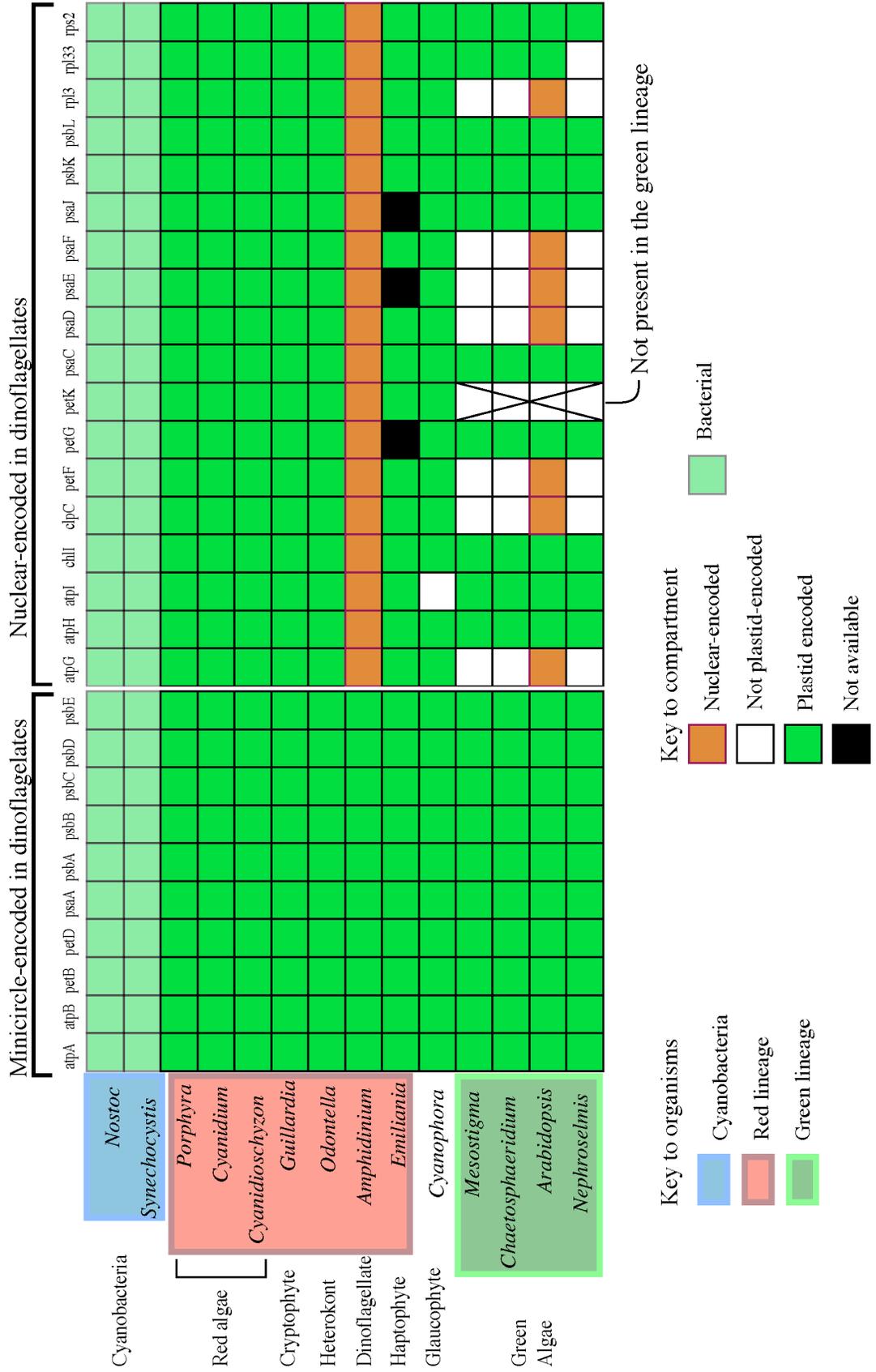


Figure III.2: Analysis of individual genes that are minicircle-encoded in dinoflagellates.

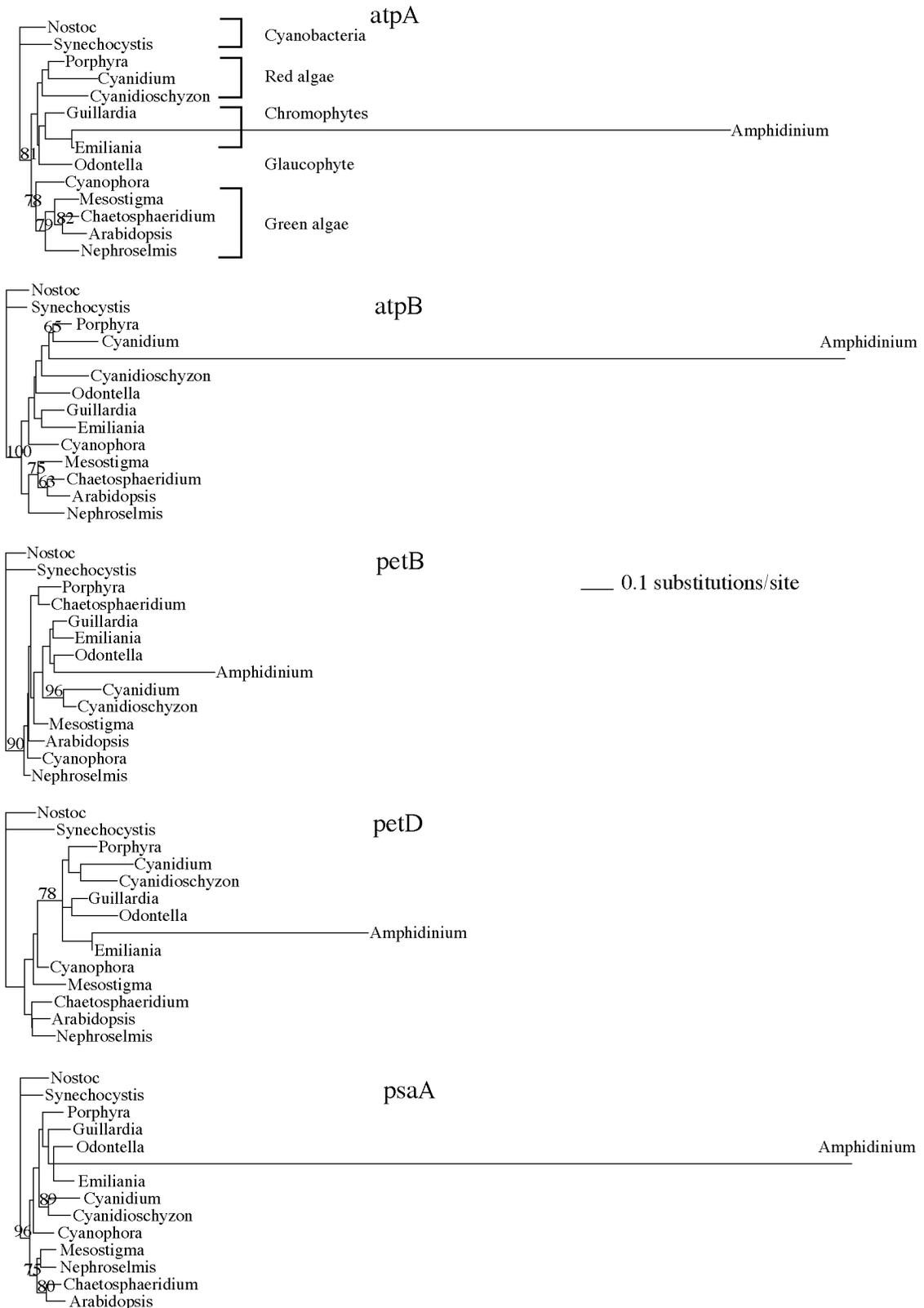
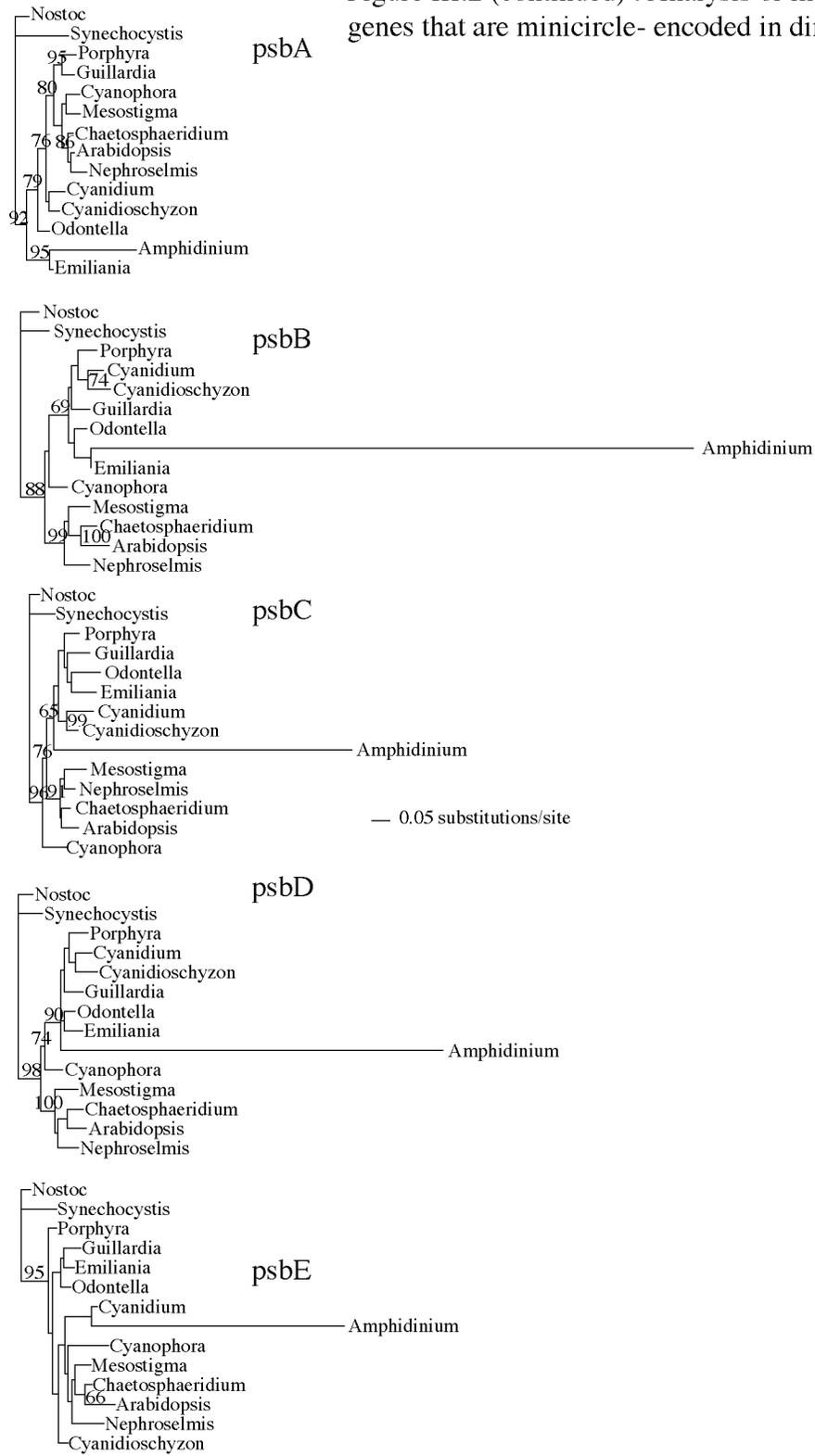


Figure III.2 (continued) : Analysis of individual genes that are minicircle- encoded in dinoflagellates.



**Table III.1 Useable alignment size and bootstrap proportions for selected groups using genes that are encoded on minicircles in dinoflagellates**

Genes were analyzed without dinoflagellates. The + indicates that the feature is present on the best tree without bootstrap support and the x indicates that the feature is absent from the best tree.

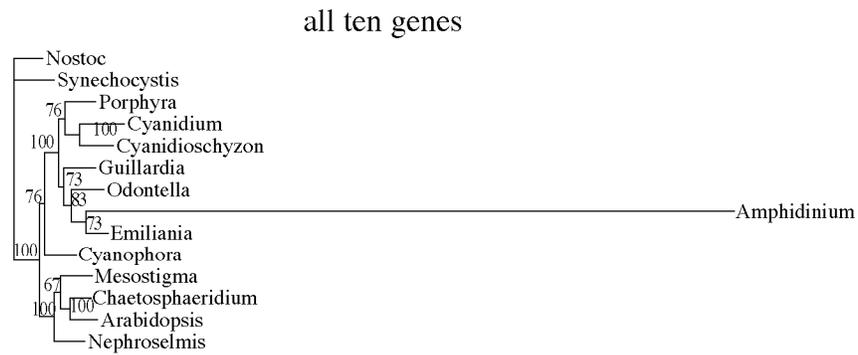
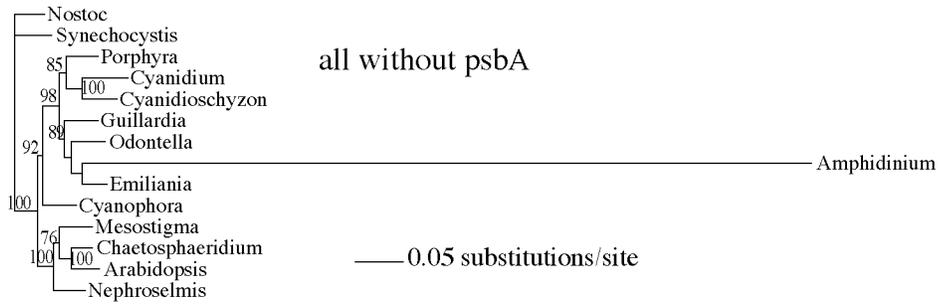
Gene	Alignment size (bases)	Cyanobacteria monophyletic	Red lineage monophyletic	Green algae monophyletic	Chromophytes monophyletic	<i>Odontella</i> + <i>Emiliana</i>
<i>atpA</i>	930	75	51	96	x	x
<i>atpB</i>	942	100	x	76	x	x
<i>petB</i>	440	99	x	x	70	x
<i>petD</i>	308	56	85	x	x	x
<i>psaA</i>	1464	100	80	99	69	73
<i>psbA</i>	661	100	x	x	x	x
<i>psbB</i>	1020	100	100	100	x	80
<i>psbC</i>	678	100	100	100	+	97
<i>psbD</i>	136	100	100	100	52	x
<i>psbE</i>	908	85	100	50	x	x
All above	7264	100	100	100	77	98
All minus <i>psbA</i>	6906	100	100	100	88	89

**Table III.2 Bootstrap proportions for selected groups using genes that are encoded on minicircles in dinoflagellates**

Genes were analyzed with dinoflagellates. The + indicates that the feature is present on the best tree without bootstrap support and the x indicates that the feature is absent from the best tree.

gene	Cyanobacteria monophyletic	Red lineage monophyletic	Green algae monophyletic	Chromophytes monophyletic	<i>Amphidinium</i> + <i>Emiliana</i>
<i>atpA</i>	81	+	79	+	+
<i>atpB</i>	100	+	+	x	x
<i>petB</i>	90	x	+	+	
<i>petD</i>	+	78	+	+	+
<i>psaA</i>	96	54	75	+	+
<i>psbA</i>	92	x	86	x	95
<i>psbB</i>	88	69	99	x	+
<i>psbC</i>	96	65	91	x	x
<i>psbD</i>	98	90	100	x	x
<i>psbE</i>	95	x	x	x	x
all above	100	100	100	73	73
minus <i>psbA</i>	100	98	100	89	58

Figure III.3: Analysis of concatenated genes that are minicircle-encoded in dinoflagellates.



The AU test can indicate if one tree (or hypothesis) is significantly better than another by comparing the site likelihoods of the best trees compatible with one hypothesis to the best tree compatible with another hypothesis. In this instance the dinoflagellate sequence was excluded from the *psbA* analysis and the best tree compatible with red lineage monophyly was compared to the best unconstrained tree (cyanobacteria + haptophyte). The best unconstrained tree had an AU value of 0.936 while the tree constraining the haptophyte to the red algal lineage had a value of 0.064. When using the *psbA* gene both the bootstrap values and the AU test indicate a strong relationship of haptophytes and cyanobacteria. In the case of the minicircle genes *psbA* was considered to be incongruent with the other genes in this dataset. When *psbA* was excluded and the remaining nine minicircle genes were concatenated the bootstrap support for chromophyte monophyly increased compared with the analysis of all the minicircle genes (Fig. III.3, Table III.2), while support for the dinoflagellate + haptophyte clade was reduced. Three other minicircle genes (*atpA*, *petD* and *psbB*) indicate a dinoflagellate + haptophyte clade while *psaA* is ambiguous and *petB* indicates a heterokont + dinoflagellate clade (Fig. III.2).

### III.2.2 Plastid genes from the dinoflagellate EST survey

This group of eighteen genes produces trees that fall into four major categories based on the relationship of the dinoflagellate. Several genes *atpI*, *petG*, *psaF*, *psaJ*, and *rpl33* (Fig. III.4) place the dinoflagellates with the red algal lineage. Other genes are ambiguous among them *clpC*, *psaC*, *psaE*, and *psbK* (Fig. III.5), where the dinoflagellate is positioned between either cyanobacteria or green algae and a member of the red

lineage. Some genes place the dinoflagellates with the cyanobacteria (*chlI*, *petK*, *psaD*, *psbL*; Fig. III.6) or with the green algae (*atpG*, *atpH*, *petF*, *rpl3* and *rps2*; Fig. III.7). When these individual genes are analyzed without the dinoflagellate sequence (trees not shown) there is bootstrap support for monophyletic cyanobacteria from seven genes, red lineage monophyly from five genes, and three genes support green lineage monophyly (Table III.3). However the remaining chromophytes are not monophyletic in any of these individual gene analyses (Table III.3).

#### *III.2.2.1 Concatenated analyses of the nuclear-encoded genes from dinoflagellates*

Concatenated analysis of the genes that are nuclear-encoded in dinoflagellates would allow congruent signal from several smaller genes to produce better bootstrap support and to discover the sister taxon to the dinoflagellate plastid. However, there is conflicting signal from different genes, so bootstrap support and the AU test were used to create congruent groups of genes for concatenated analyses.

Three genes were rejected because of bootstrap support for a dinoflagellate + cyanobacteria (*petK*) or a dinoflagellate + green (*petF*, *atpG*) clade. The AU test was used here to indicate which other genes rejected dinoflagellate + red lineage monophyly (Table III.4). In this instance the best unconstrained tree was compared with the best tree that placed the dinoflagellate in a monophyletic red lineage, cyanobacterial lineage or green lineage.

Two levels of significance were used, the first level was if the dinoflagellate + red lineage tree had an AU value of  $<0.1$ . In a general sense this value suggests that the dinoflagellate + red lineage would be found in only 10% of trees generated from these data. The genes *atpH*, *chlI*, *petF* and *psbK* strongly rejected red lineage + dinoflagellate

monophyly (AU value < 0.1). The remaining genes are *atpI*, *clpC*, *petG*, *psaC*, *psaD*, *psaE*, *psaF*, *psaJ*, *psbL*, *rpl3*, *rpl33*, *rps2* (2750 bases). This dataset is referred to as the red stringent dataset. When the cutoff AU value is < 0.2 then *psaD*, *psaE*, *psbL* and *rpl33* are rejected leaving *atpI*, *clpC*, *petG*, *psaC*, *psaF*, *psaJ*, *rpl3* and *rps2* (2202 bp), and this dataset is called red relaxed. This produces two concatenated datasets that are increasingly exclusive of genes.

Analyses of these two concatenated datasets (derived from genes that are nuclear-encoded in dinoflagellates) produced trees that place the dinoflagellate with *Odontella* with 79% (red stringent dataset) and 94% (red relaxed dataset) bootstrap support (Fig. III.8). The branching pattern shown in these trees is exactly the same when dinoflagellates are excluded, with *Odontella* at the base of the red lineage and *Emiliana* grouped with *Porphyra*.

### 2.3 Comparison of the minicircle dataset with the nuclear-encoded dataset

Since the signal from the nuclear-encoded dataset and the minicircle dataset is very different more AU tests were performed using more specific hypotheses. To compare support for monophyly of chromists, and the sister taxon to dinoflagellate the AU test was performed comparing the *Amphidinium* + *Odontella* with the *Amphidinium* + *Emiliana* hypotheses, as well as chromophyte monophyly (Table III.5). These data highlight the conflict between the minicircle and nuclear-encoded genes in dinoflagellates, with the minicircle data strongly rejecting an *Amphidinium* + *Odontella* hypothesis and the nuclear-encoded data even more strongly rejecting an *Amphidinium* + *Emiliana* hypotheses (Table III.5).

Figure III.4: Trees using plastid genes that have been transferred to the dinoflagellate nucleus. These genes group the dinoflagellate with members of the red plastid lineage.

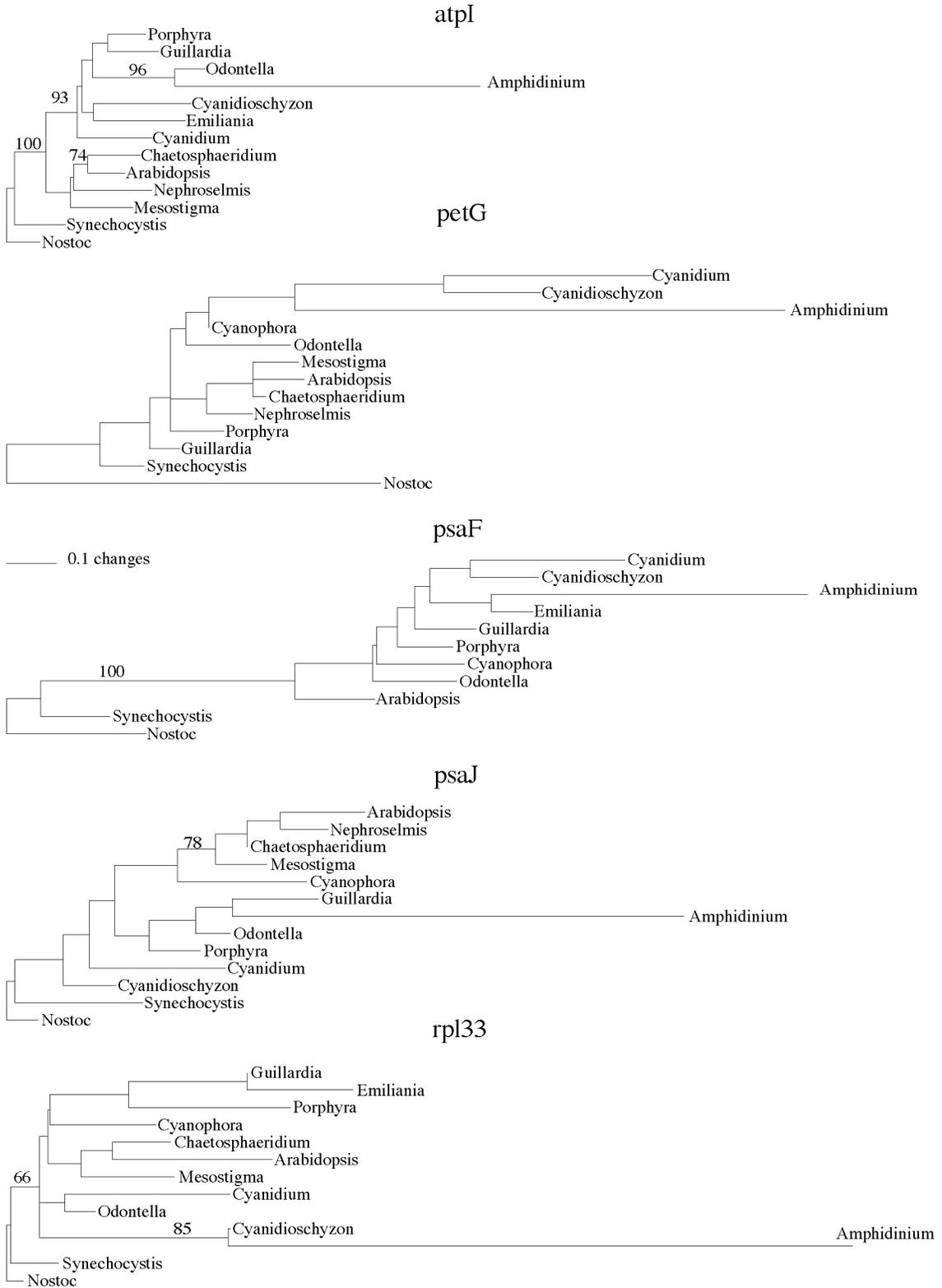


Figure III.5: Trees using plastid genes that have been transferred to the dinoflagellate nucleus. These genes group the dinoflagellate between cyanobacteria, green algae or the red algal lineage without support.

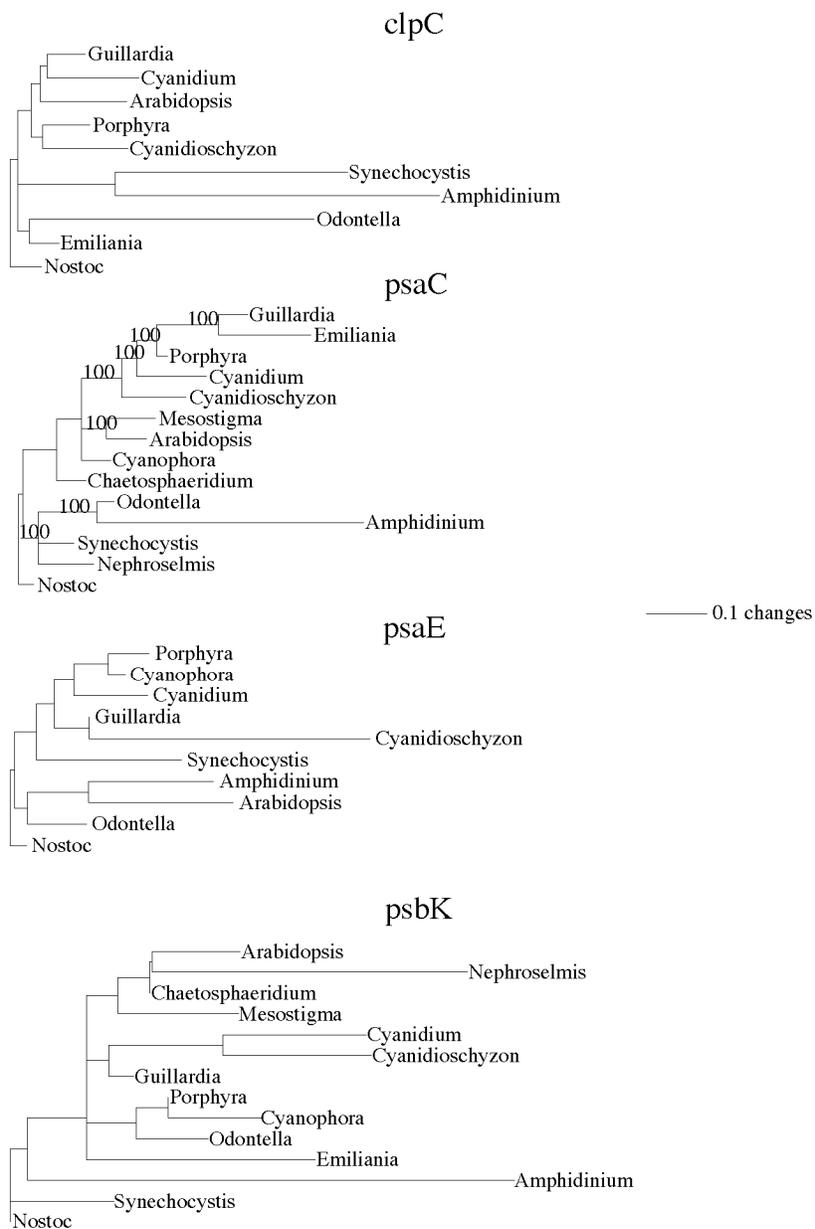


Figure III.6: Trees using plastid genes that have been transferred to the dinoflagellate nucleus. These genes group the dinoflagellate within the cyanobacteria or sister to the cyanobacteria with support.

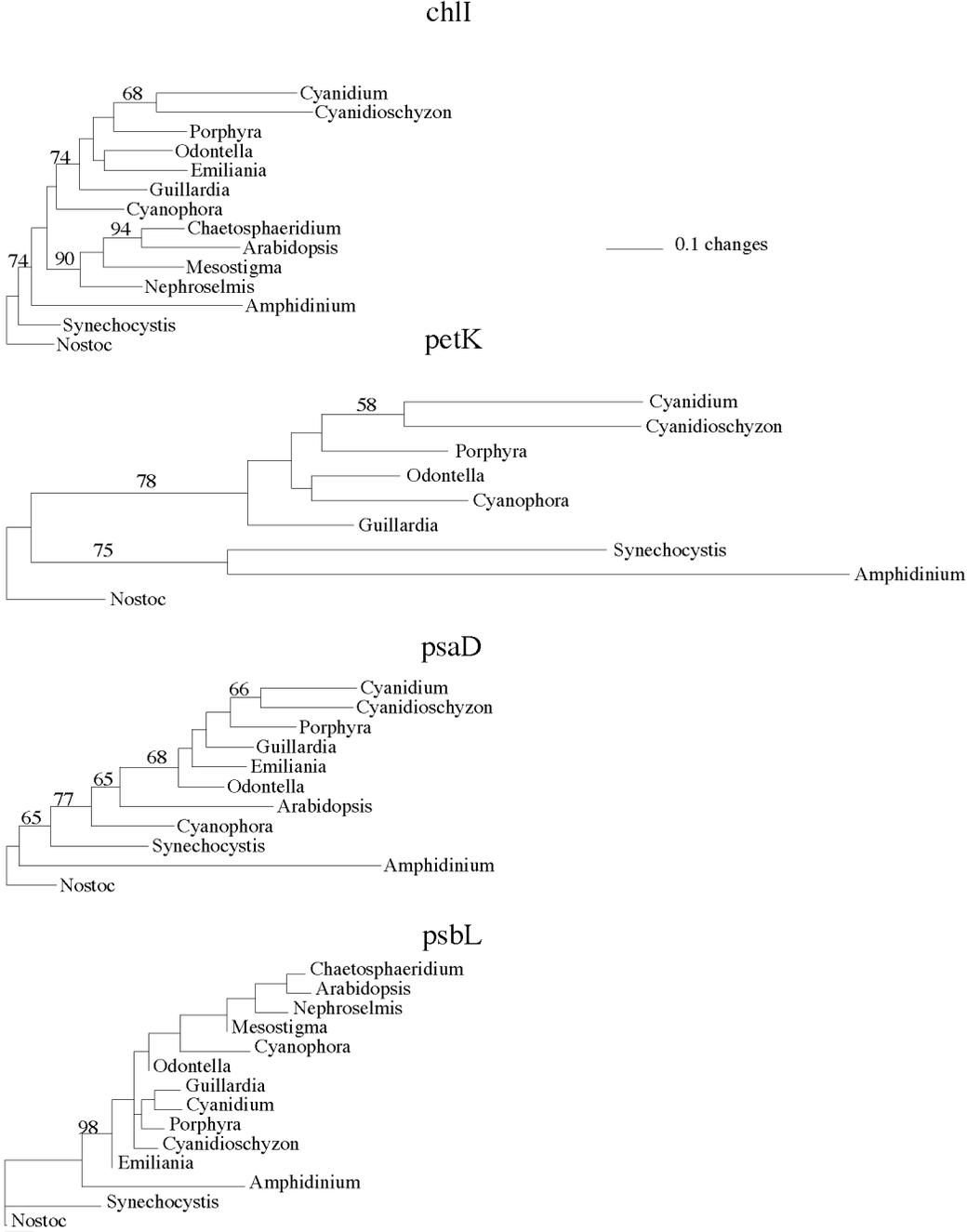
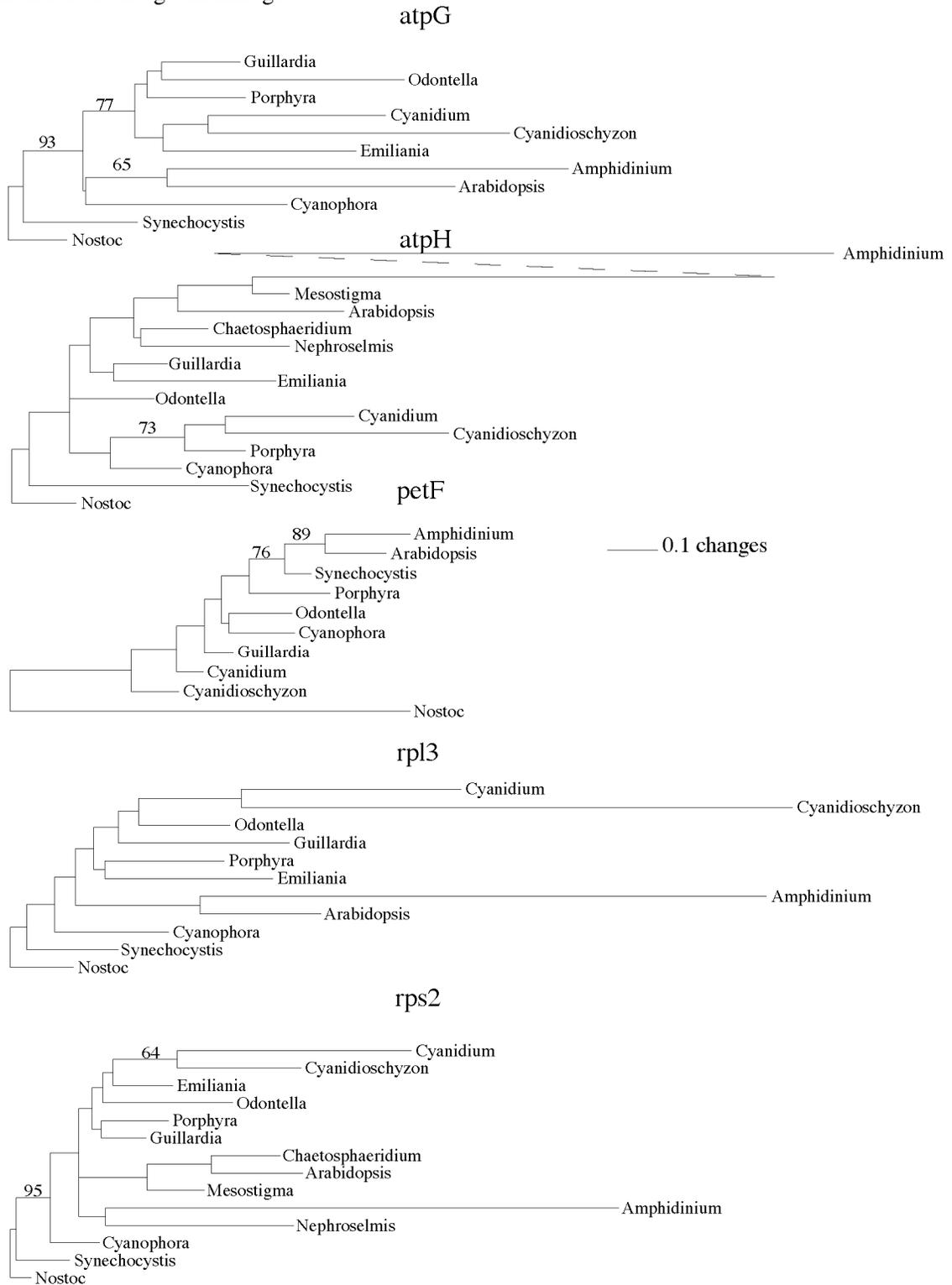


Figure III.7: Trees using plastid genes that have been transferred to the dinoflagellate nucleus. These genes group the dinoflagellate with or within members of the green lineage.



**Table III.3 Bootstrap support and size of useable alignment using the set of plastid genes that are nuclear-encoded in dinoflagellates.**

In this analysis dinoflagellates are excluded. The + indicates that the feature is present on the best tree without bootstrap support and the x indicates that the feature is absent from the best tree.

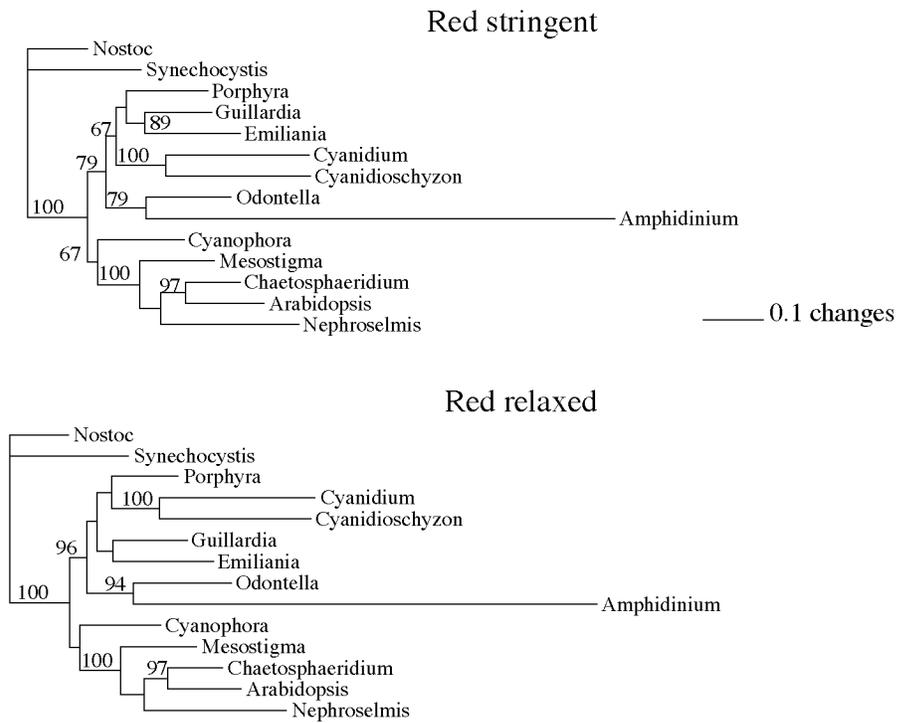
gene	Useable size (bases)	Cyanobacteria monophyletic	Red lineage monophyletic	Green algae monophyletic	Chromophytes monophyletic
<i>atpG</i>	274	87	72	NA	x
<i>atpH</i>	158	+	x	58	x
<i>atpI</i>	454	98	95	63	x
<i>chlI</i>	688	90	84	96	x
<i>clpC</i>	300	x	x	x	x
<i>petF</i>	198	x	x	x	x
<i>petG</i>	77	x	x	+	x
<i>petK</i>	278	95	x	NA	x
<i>psaC</i>	168	x	x	x	x
<i>psaD</i>	278	64	79	NA	x
<i>psaE</i>	122	x	x	x	NA
<i>psaF</i>	242	100	77	NA	x
<i>psaJ</i>	68	52	x	54	x
<i>psbK</i>	82	+	x	x	x
<i>psbL</i>	78	93	x	56	x
<i>rpl3</i>	454	x	+	NA	x
<i>rpl33</i>	136	62	x	+	x
<i>rps2</i>	440	99	+	80	x
All	4494	100	98	100	x

**Table III.4 AU test values for three different hypotheses when dinoflagellates are included.**

Initially genes that had bootstrap support for a non-red ancestry of the dinoflagellate gene were rejected from further analyses. Then genes that rejected a monophyletic red lineage with dinoflagellates with an AU value <0.2 were rejected.

gene	best tree	<i>Amphidinium</i> + red	<i>Amphidinium</i> + cyanobacteria	<i>Amphidinium</i> + green	Genes rejected by bootstrap	<i>Amphidinium</i> + red < 0.1	<i>Amphidinium</i> + red < 0.2
<i>atpG</i>	0.723	0.282	0.283	0.723	Yes		
<i>atpH</i>	0.640	<b>0.071</b>	0.514	0.437		Yes	Yes
<i>atpI</i>	0.963	0.963	0.048	0.047			
<i>chlI</i>	0.781	<b>0.098</b>	0.781	0.331		Yes	Yes
<i>clpC</i>	0.636	0.587	0.086	0.135			
<i>petF</i>	0.937	<b>0.041</b>	0.051	0.937	Yes	Yes	Yes
<i>petG</i>	0.626	0.336	0.491	0.378			
<i>petK</i>	0.877	<b>0.133</b>	0.877	NA	Yes		Yes
<i>psaC</i>	0.573	0.425	0.554	0.156			
<i>psaD</i>	0.882	<b>0.118</b>	0.882	0.118			
<i>psaE</i>	0.884	<b>0.200</b>	0.227	0.884			Yes
<i>psaF</i>	0.750	0.548	0.236	0.219			
<i>psaJ</i>	0.834	0.330	0.287	0.100			
<i>psbK</i>	0.542	<b>0.070</b>	0.705	0.339		Yes	Yes
<i>psbL</i>	0.820	<b>0.189</b>	0.820	0.288			Yes
<i>rpl3</i>	0.565	0.394	0.004	0.507			
<i>rpl33</i>	0.739	<b>0.136</b>	0.421	0.383			Yes
<i>rps2</i>	0.592	0.570	0.192	0.592			

Figure III.8: Trees constructed using plastid genes that have been transferred to the dinoflagellate nucleus. This is an analysis of concatenated datasets. The first tree is when a more stringent criterion was used to exclude incompatible genes (less genes were excluded). The second is when a less stringent criterion was used to exclude incompatible genes (more genes were excluded).



**Table III.5 AU test values for selected hypotheses.**

In cases where the best tree is compatible with one of the constraints it will have the same value as that constraint.

gene	best tree	<i>Amphidinium</i> + red	<i>Amphidinium</i> + cyanobacteria	<i>Amphidinium</i> + Green	<i>Amphidinium</i> + <i>Emiliana</i>	chromophyte monophyly	<i>Amphidinium</i> + <i>Odontella</i>
<i>atpA</i>	0.951	0.951	0.085	0.106	0.951	0.951	0.045
<i>atpB</i>	0.104	0.294	0.639	0.460	0.567	0.445	0.392
<i>petB</i>	0.605	0.402	0.425	0.120	0.592	0.605	0.605
<i>petD</i>	0.700	0.700	0.006	0.180	0.686	0.608	0.201
<i>psaA</i>	0.752	0.752	0.377	0.165	0.380	0.752	0.377
<i>psbA</i>	0.977	0.031	0.078	0.002	0.977	0.001	0.004
<i>psbB</i>	0.589	0.589	0.512	0.035	0.589	0.459	0.302
<i>psbC</i>	0.800	0.800	0.043	0.437	0.178	0.451	0.273
<i>psbD</i>	0.368	0.769	0.257	0.123	0.260	0.497	0.384
<i>psbE</i>	0.687	0.450	0.073	0.222	0.378	0.378	0.146
all above	0.674	0.470	0.076	0.187	0.395	0.395	0.118
minus <i>psbA</i>	0.962	0.962	0.007	0.007	0.962	0.962	0.063
Red tight	0.897	0.897	0.248	0.170	0.002	0.177	0.897
Red relaxed	0.958	0.958	0.025	0.053	4.00x10 <sup>-04</sup>	0.114	0.958

### III.4.3. Discussion

#### 3.1 Assessing monophyly of chromophyte plastids

Although the hypothesis that chlorophyll *c* containing plastids are monophyletic is widely accepted (Baldauf 2003; Palmer 2003) there is little comprehensive data supporting this assertion largely because of the lack of tractable data from dinoflagellates. The minicircle data presented here suggests that the chlorophyll *c* containing plastids are monophyletic when all four chromist lineages are present, although there are significant issues of saturation and long branch problems with the dinoflagellate. Data gleaned from an EST survey of *Amphidinium* revealed large scale plastid to nucleus gene transfer (Bachvaroff et al. 2004) and the analysis of these genes reveals significant compatibility issues, and does not indicate chromophyte monophyly.

#### 3.1 Minicircle data

The minicircle data support a relationship of haptophyte and dinoflagellate plastids. However, the remarkable branch lengths leading to the dinoflagellate suggest that saturation is likely. There is support for a dinoflagellate + haptophyte clade using concatenated *psaA* and *psbA* sequences (Yoon et al. 2002; Shalchian-Tabrizi 2003), although again the branch lengths are extreme. The strong signal supporting this relationship from the *psbA* sequence may indeed reflect plastid phylogeny, but this gene groups the haptophyte + dinoflagellate clade with cyanobacteria. One reason for the different results in the present study may be that in this case full-length sequences are used, whereas the previous study used partial *psbA* PCR amplicons (Yoon et al. 2002). An analysis of the *psbB* gene with better taxon sampling than the present study also

supports the haptophyte + dinoflagellate relationship (Chapter V). These genes are perhaps the most genuine for phylogenetic analysis of plastids since they are readily aligned and found in the chloroplast genome. Unfortunately in dinoflagellates genes encoded on minicircles have evolved with an extremely different rate compared to genes encoded on typical plastid genomes and it is difficult to believe any relationship that they support (see Chapter IV for further discussion of rate heterogeneity).

### 3.2 Genes that are nuclear-encoded in dinoflagellates

The eighteen genes in this category have been transferred from the plastid to the nucleus in dinoflagellates (Bachvaroff et al. 2004), and in some cases in the green lineage although in the red lineage they are plastid-encoded (Fig. III.1). Useable alignments of these genes are generally shorter (Table III.3), and when dinoflagellates are excluded the support for known relationships (i.e. red lineage, green and cyanobacterial monophyly) is not as strong as the minicircle dataset (Table III.1 vs. Table III.3). When the dinoflagellates are included almost all of the known relationships have no bootstrap support, and the dinoflagellates are placed with the red (Fig. III.4), cyanobacterial (Fig. III.6) and green (Fig. III.7) lineage depending on which gene is used. For example the *atpG* gene supports a relationship of the dinoflagellate with *Arabidopsis*, but in this case both the dinoflagellate and *Arabidopsis* genes are encoded in the nucleus. Is this phylogenetic signal derived from convergence between otherwise unrelated genes when they are transferred to the nucleus? Or is there any phylogenetic signal given the long branches leading to the dinoflagellate? A similar pattern might be true for the cyanobacterial relationship implied by *petK*. While other eukaryotes retain these genes in the plastid dinoflagellates have moved them to the nucleus. This may produce an artifact

whereby the dinoflagellate nuclear genes are pushed into the outgroup because of shared selection for specific amino acids or nucleotides in this compartment.

An alternate explanation for these results is that the dinoflagellate plastid is the result of a mosaic pattern of evolution, with dinoflagellates taking genes from diverse lineages and incorporating them into their suite of nuclear-encoded plastid targeted genes. There is no doubt that these genes are expressed since they were recovered from an EST survey, and it is likely that the most highly expressed versions of these genes were sampled. However, dinoflagellates would have had to lose (or fail to gain) the genuine red lineage version of each gene, and then somehow acquire this same gene from another lineage. In any other group of organisms this would be easily dismissed, but dinoflagellates are remarkably bizarre. This very type of transfer has been documented for the anomalous rubisco that dinoflagellates use in their plastids (Morse et al. 1995; Rowan et al. 1996; Jenks and Gibbs 2000). Furthermore, dinoflagellates themselves have acquired endosymbionts from almost every major lineage including heterokonts (Chesnick et al. 1996), cryptophytes (Takishita et al. 2002), haptophytes (Tengs et al. 2000), and green algae (Watanabe et al. 1990).

### *3.2.1 Concatenated datasets*

Analyses of individual genes that are nuclear-encoded in dinoflagellates revealed strongly incongruent results. Therefore, the genes were sorted into compatible categories and then concatenated. Genes that either had bootstrap support for a non-red ancestry or strongly rejected a red ancestry were excluded (Table III.4). The logic behind the exclusion of selected data seems circular, with data that rejects the favored hypothesis being excluded so that the favored hypotheses can be more strongly supported. However,

genes that have incompatible signal are unlikely to produce a good tree. Secondly, there is little doubt that the dinoflagellate plastid (or at least most of its genes) came from the red algal lineage. The hypothesis that these concatenated data are testing is the more specific question of whether the dinoflagellate plastid is more closely related to the haptophyte or to the heterokont plastid. Additionally if, for example, dinoflagellate *petF* and *chlI*, are actually derived from the green and cyanobacterial lineage, then there would be no reason to include them.

The genes that are nuclear-encoded in dinoflagellates were concatenated into compatible groups and analyzed. The signal from both of the concatenated datasets (one being more stringent than the other) is that the dinoflagellate plastid is more closely related to the heterokont plastid. The support for this relationship goes up when more data is excluded (Fig. III.8). However, these concatenated datasets do not recover a monophyletic chromophyte lineage, either with (Fig. III.8) or without (Table III.3) dinoflagellates. The conflict between the datasets is therefore not just a property of including the dinoflagellates in the analysis. In any case the chromophytes are not separated from each other by nodes with strong bootstrap support using these genes.

### *3.3 Implications and models of chloroplast evolution*

The single origin of the chlorophyll *c* containing plastid from red algae (Yoon et al. 2002; Harper and Keeling 2003) has been assumed to correspond to host phylogeny and has been called the chromalveolate hypothesis (Fast et al. 2001). However, a single origin of the chromophyte plastid from red algae is not incongruent with multiple endosymbiotic events. One model of chloroplast evolution is that the chlorophyll *c*

containing plastid was passed among cryptophytes, dinoflagellates, haptophytes and heterokonts. This model requires multiple endosymbiotic events after the first plastid acquisition from the red algae and could produce a tree like Figure III.3. For example, after cryptophytes acquired their plastid from a red alga, cryptophytes themselves could have been engulfed by a heterokont, haptophyte or dinoflagellate (Cavalier-Smith et al. 1994). This model is completely compatible with plastid trees showing chromophyte monophyly and with the GAPDH data. The cyanobacterial GAPDH gene has been replaced by a cytosolic version in all chromophyte lineages (Fast et al. 2001). If this substitution event had already occurred in the cryptophyte before it was engulfed then this gene would be passed on to the next host just like all of the other nuclear-encoded plastid-targeted genes.

If the chromophyte plastid was passed from one host to another then it might be difficult to discern between separate acquisitions from one lineage or passing the plastid in series through lineages. In other words, the plastids that dinoflagellates, and haptophytes possess may have arisen separately from one lineage of heterokonts. Alternatively dinoflagellates could have acquired a plastid from haptophytes (who in turn had acquired it from heterokonts) and the phylogenies would be remarkably similar. The difficulty in discerning between these scenarios is reflected by the plastid gene trees. When the incongruent *psbA* gene was removed from the minicircle dataset the support for chromophyte monophyly was increased while support for a specific branching order within the chromophytes decreased. As a comparison, additional taxon sampling embeds a monophyletic chromophyte lineage (minus dinoflagellates) within a broader red algae (as is seen in Yoon et al. 2002b). To embed individual chromist lineages within another

chromist lineage would require finer taxon sampling and powerful data. A tree that showed dinoflagellate and haptophyte plastids as sister taxa embedded within a broader heterokont lineage might be the outcome. Certainly the development of stronger chloroplast phylogenies should help to test specific elements of the chromalveolate hypothesis.

### **III.5.4. Methods**

#### *4.1 Sequence acquisition*

To obtain partial chloroplast genome data from the haptophyte *Emiliana huxleyi* (CCMP 373) the algae were 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 14hr/10hr L:D cycle at 24  $\mu\text{mol photons/m}^2\cdot\text{s}$ . Total DNA was isolated from 8L of culture using CTAB followed by chloroform extraction. Isopycnic ultracentrifugation was used to separate mitochondrial chloroplast and nuclear DNA fractions (Chesnick and Cattolico 1993). DNA from the least dense fraction corresponded to the mitochondrial genome, and the intermediate density fraction corresponded to the chloroplast genome. This DNA was digested with *HindIII* and ligated to pGEM®-3Zf(+). Plasmid DNA from arbitrarily selected clones was isolated using the miniprep method (Sambrook et al. 1989) sequenced and subjected to blast analysis for putative identification.

#### **4.1.1 Sequencing**

Sequencing reactions were performed on an ABI 3100 using the Big-Dye Terminator kit from ABI. The sequencing reaction volumes were reduced according to the manufacturer's directions for 384 well plates. Sequences were edited and contiguous

sequences were assembled using Sequencher. Primer walking was used to completely sequence selected genes.

#### 4.2 Alignments

Nucleotide alignments for the 28 genes were made by downloading the appropriate sequences from GenBank as well as from the dinoflagellate EST database (<http://oxrid.umd.edu>). GenBank numbers are as follows *Arabidopsis thaliana* (NC\_000932; Sato et al. 1999), *Chaetosphaeridium globosum* (NC\_004115; Turmel et al. 2002), *Cyanidium caldarium* (Glockner et al. 2000), *Cyanidioschyzon merolae* (NC\_004799; Ohta et al. 2003), *Cyanophora paradoxa* (NC\_001675; Stirewalt et al. 1995), *Emiliana huxleyi* (unpublished), *Guillardia theta* (NC\_000926; Douglas and Penny 1999), *Mesostigma viride* (NC\_002186; Lemieux et al. 2000), *Nephroselmis olivacea* (NC\_000927; Turmel et al. 1999), *Nostoc sp.*PCC7120 (NC\_003272; Kaneko et al. 2001), *Odontella sinensis* (NC\_001713; Kowallik et al. 1995), *Porphyra purpurea* (NC\_000925; Reith and Munholland 1995), and *Synechocystis sp.* PCC6803 (NC\_000911; Kaneko et al. 2001). The minicircle genes used are described from *Amphidinium carterae* and *Amphidinium operculatum*, although the two strains appear to be identical in their minicircle gene sequences (Barbrook and Howe 2000; Hiller 2001; Bachvaroff et al. 2004). The sequences were imported into MacClade 4.0 (Maddison and Maddison 2000) and in most cases were easily aligned by eye. However, genes were also translated and aligned using ClustalW as a guide to the nucleotide alignment.

##### 4.2.1 Analysis

For all single and multi-gene analyses PAUP\*4b10 (Swofford 2002) was used and the third codon position was excluded from all analyses. First a Fitch-Margoliash

tree was constructed using LogDet distances, then Maximum Likelihood (ML) parameters for the General Time Reversible model of evolution with Invariant sites and gamma correction (4 categories) were estimated from this tree. These parameter estimates were used in the ML heuristic search repeated three times with different random addition order. For bootstrapping a single heuristic search with full branch swapping (TBR) was used.

To perform the AU test constraint trees were constructed according to the hypothesis being tested (Shimodaira 2000; Shimodaira and Hasegawa 2001). Trees that constrained the dinoflagellates to a monophyletic red algal lineage, to a monophyletic green lineage or to a monophyletic cyanobacterial lineage were constructed using the same method described above. Site likelihoods for these trees as well as the most likely tree in unconstrained analyses were exported from PAUP. The CONSEL package (Shimodaira and Hasegawa 2001) was then used to test the significance of different trees. For some genes and gene combinations additional hypotheses were tested. These additional hypotheses were dinoflagellate and haptophyte monophyly, dinoflagellate and heterokont monophyly and chromophyte monophyly. These hypotheses were tested along with the three previous hypotheses in the same way as described above using the site likelihoods of the most likely tree corresponding to the hypothesis.

## **Chapter IV Evolutionary rates of dinoflagellate chloroplast genes: heterogeneous rates of evolution between plastid-encoded and plastid-targeted genes**

### **IV.1. Abstract**

Peridinin-pigmented dinoflagellates are unusual in that their plastid-encoded genes are found on minicircles and based on phylogenetic trees these genes appear to be rapidly evolving. Additionally, most typically plastid-encoded genes have been transferred to the nucleus. The evolutionary rate of minicircle genes was compared to the rate of genes transferred from the chloroplast to the nucleus. The almost complete chloroplast data from the haptophyte *Emiliana huxleyi* as well as EST data from the peridinin dinoflagellate *Amphidinium carterae* allows for direct comparison of the dinoflagellate with other organisms with secondary plastids. For comparison, two plastid-associated genes, GAPDH and psbO, that are unlikely to have been plastid-encoded at the time of endosymbiosis were also tested. Based on both distance-based relative rate tests as well as likelihood ratio tests dinoflagellate minicircle and formerly plastid-encoded genes have unusually high rates of evolution. The plastid-associated genes that were probably encoded in the nucleus of the previous host have rates of evolution that are similar to other taxa, suggesting that the accelerated evolution is specific to genes that were likely plastid-encoded at the time of endosymbiosis. Although there are several plausible explanations for this pattern it is likely that this acceleration in rate has overcome the selective advantages of retaining genes in the plastid and promoted plastid to nucleus gene transfer.

## IV.2.Introduction

Horizontal gene transfer is an important evolutionary phenomenon involving the movement of genes within and between organisms. Endosymbiotic organelles create an opportunity to observe gene transfer and substitution both from within and outside of the cell (Gray and Spencer 1996; Martin et al. 1998; Palmer and Delwiche 1998; Delwiche 1999; Martin et al. 2002). Because dinoflagellates have secondary plastids derived from an endosymbiotic event between a dinoflagellate and another eukaryote a different pattern of gene transfer is required to integrate the endosymbiont than in primary endosymbiosis. The peridinin-containing dinoflagellates are particularly interesting because their plastids appear to evolve under different constraints than other plastids. Dinoflagellate plastid-encoded genes have long branches on phylogenetic trees suggesting a high mutation rate in these genes (Takishita and Uchida 1999; Zhang et al. 1999; Zhang et al. 2000). However, the evolutionary rate of dinoflagellate plastid-encoded genes has not been tested using relative rate tests.

Dinoflagellates appear to have retained a small number of rapidly-evolving genes in their plastids and these genes are encoded in single or double gene minicircles (Zhang et al. 1999). In the case of *Amphidinium* eleven protein-coding minicircle genes are known (Barbrook and Howe 2000; Barbrook et al. 2001; Hiller 2001) and many genes that would typically be encoded in the chloroplast genome are found in the nuclear genome (Bachvaroff et al. 2004). A similar pattern of transfer was also found in the peridinin dinoflagellate *Alexandrium tamarense* (Hackett et al. 2004). This atypical pattern of gene transfer and evolution allows for direct comparison of the few genes that

are retained on the plastid genome with genes that have migrated from the plastid to the nucleus.

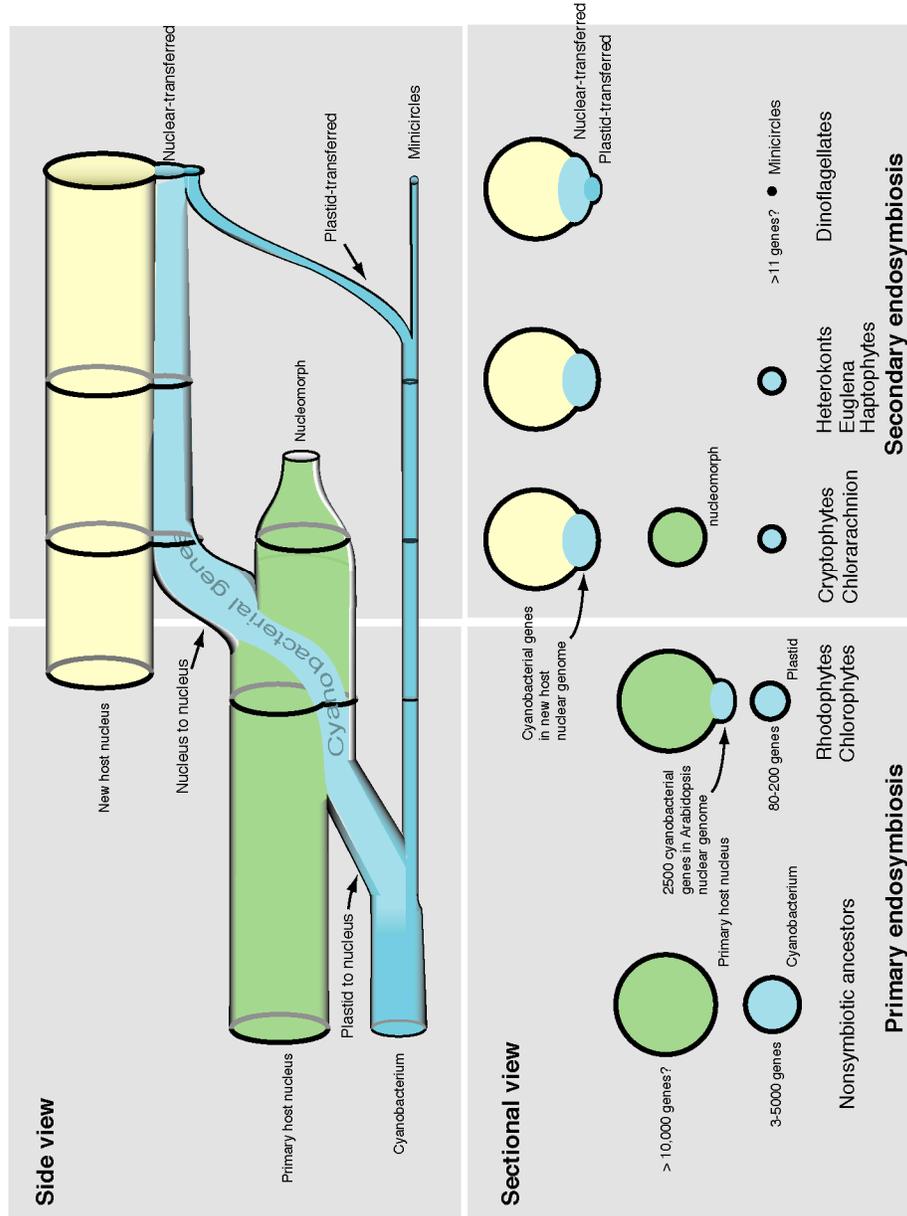
Three classes of plastid-associated genes can be compared to make inferences about the evolution of plastid-associated genes. Here plastid-associated genes are those that are encoded in, targeted to, or derived from the plastid. The first class of genes is found on minicircles in dinoflagellates and these genes are likely to be located in the chloroplast (Takashita et al. 2003). The second class of genes are those that were probably plastid-encoded at the time of endosymbiosis, but subsequently were transferred from the plastid to the nuclear genome in dinoflagellates. These genes will henceforth be called “plastid-transferred genes”. Although the genome of the ancestral plastid cannot be known with certainty, the dinoflagellate plastid is likely derived from the red plastid lineage and all other red lineage plastids retain a common set of genes in their chloroplast genomes (Kowallik et al. 1995; Reith and Munholland 1995; Martin et al. 1998; Sanchez-Puerta et al. 2004). It is therefore a good inference that these genes were in the plastid genome of the previous chloroplast host. The third class of genes referred to here as, “nuclear-transferred”, is unique to secondary endosymbiosis; these genes are plastid-associated genes transferred from the nucleus of the previous host to the nucleus of the new host. As with the plastid-transferred genes this inference is again based on known chloroplast genomes. There are thought to be two thousand or more proteins required for plastid maintenance, yet the richest of plastid genomes known contain less than three hundred protein coding genes (Arabidopsis Genome Initiative 2000; Martin et al. 2002). The remaining genes are nuclear-encoded in known eukaryotes both from the red and green plastid lineages. Dinoflagellates are likely to have acquired their plastid from a

eukaryote that had already transferred most of its plastid-associated genes to its own nucleus. Although the diverse biochemical functions of the dinoflagellate chloroplast remain largely uncharacterized it is clear that chloroplast-encoded genes constitute only a tiny fraction of the required genes. These nuclear-encoded plastid-targeted genes would have to be transferred from the previous host nucleus to the dinoflagellate host nucleus to maintain the plastid in its new host.

The proposed pattern of gene transfer in primary and secondary endosymbiosis based on the model described above is summarized in figure IV.1. After the primary endosymbiotic event between a cyanobacterium and a eukaryote some cyanobacterial genes were transferred to the host nucleus, while others were either lost or substituted by host genes. In secondary endosymbiosis these nuclear-encoded and plastid-targeted genes would have been transferred from the nucleus of the primary host to the nucleus of the secondary host. An intermediate state is found in cryptomonads and chlorarachniophytes, where a residual highly reduced nuclear genome is retained in the endosymbiont. Finally in dinoflagellates an additional level of gene transfer between the plastid and nucleus has occurred.

We have measured the evolutionary rate of plastid-associated genes that could be tentatively assigned to each of these three classes: nuclear-transferred, plastid-transferred and plastid-encoded genes. Distance based relative rate tests (Sarich and Wilson 1973) as well as likelihood ratio tests (Huelsenbeck and Rannala 1997) showed substantial differences in rate within and between these three classes of genes; this result suggests that there is a strong incongruence in the evolutionary rate of different classes of plastid-associated genes.

Figure IV.1 Gene transfer in primary and secondary endosymbiosis. The top portion of the figure shows the flow of cyanobacterial genes from chloroplast to nucleus in primary endosymbiosis, and from nucleus to nucleus in secondary endosymbiosis. The bottom portion shows sectional views through the different compartments at different stages of endosymbiosis. This diagram is schematic only, and is not meant to describe all gene transfer, just the major flow of genes from host to host nor should this diagram be interpreted as an evolutionary sequence leading to dinoflagellates.



### IV.3.Results

#### *Relative rate tests*

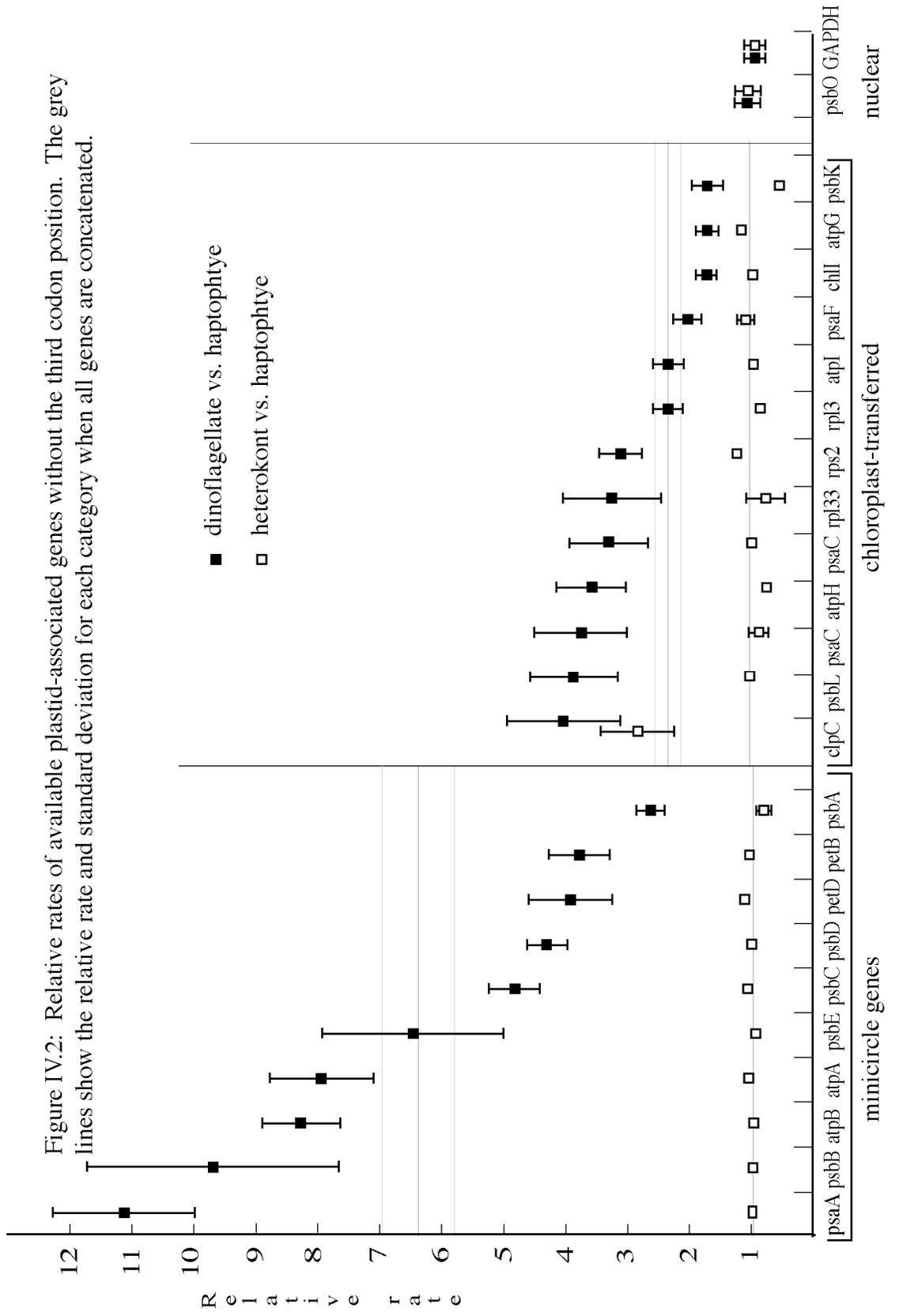
Relative rate tests use pairwise differences to determine if the evolutionary rate of a group or taxon is higher or lower than the outgroup. To compare the relative rate of dinoflagellates with the relative rate of haptophytes the pairwise distance between an outgroup taxon (e.g. *Arabidopsis*) and the dinoflagellate is divided by the pairwise distance between the same outgroup taxon and the haptophyte. When the distance between the dinoflagellate and *Arabidopsis* is similar to the distance between the haptophyte and *Arabidopsis* then the value will be close to one. If the dinoflagellate has a higher relative rate then the ratio of pairwise distances will be proportionately higher. These comparisons are then made across all outgroup pairwise comparisons and the average (mean) value as well as the standard deviation can be calculated. These distance measures are dependant on the model of sequence evolution selected (here the parameter rich GTR+I+ $\Gamma$ 4 model is used), but the comparisons are independent of a tree. The relative rate of haptophytes vs. heterokonts, organisms that also have secondary plastids from the red lineage, was compared to the relative rate of dinoflagellates vs. haptophytes.

The results of the relative rate tests are shown both in figure IV.2. In these tests the third codon position was removed from the alignment to avoid problems with saturation. Generally, the relative rate of haptophytes vs. heterokonts was close to one with the exception of *clpC* where the relative rate approaches the dinoflagellate vs. haptophyte. The relative rate of the dinoflagellate vs. haptophyte shows a clear acceleration of the rate in dinoflagellates in both plastid-encoded and plastid-transferred

genes. Although there are substantial differences in relative rate within the minicircle genes it is unclear whether these values are reflecting complete saturation. Comparing the dinoflagellate minicircle genes with plastid-transferred genes shows that in general the relative rate of minicircle genes is higher than the relative rate of plastid-transferred genes, although *clpC*, *psbL*, *psaC*, *atpH*, *rpl33* and *rps2* have a rate comparable to minicircle genes. Both of these classes, plastid-transferred and minicircle genes have a much higher rate than nuclear-transferred genes such as *psbO* and GAPDH. In both GAPDH and *psbO* the relative rate of dinoflagellates vs. haptophytes is similar to the relative rate of haptophytes vs. heterokonts.

#### *Likelihood ratio tests*

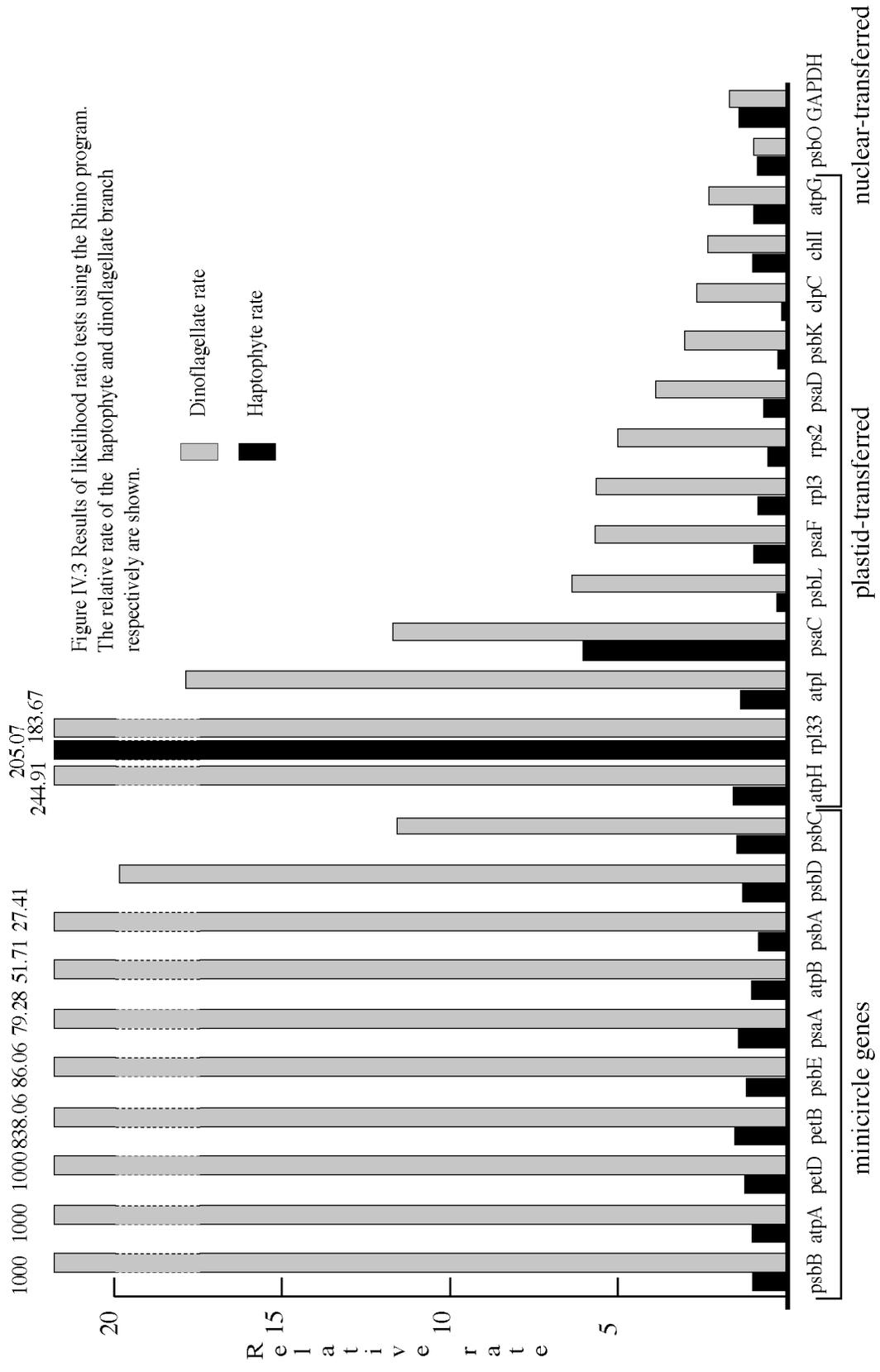
Likelihood ratio tests are used to test whether the observed difference in likelihood between two nested models is significant. In this instance the two models that were used are a single-rate parameter model compared to a two-rate parameter model. In the two-rate parameter model an additional rate parameter is added to the branch or branches of interest. If a single-rate parameter gives a tree with a significantly worse likelihood score than a tree with two-rate parameters then different rates of evolution better describe the data than a single rate of evolution. If the difference in likelihood is small and the addition of another rate parameter is not significant then the assumption of a single rate is not violated and two rates do not describe the data better than a single rate.



The same tests were applied to haptophytes (or the haptophyte lineage when applicable) to test for rate heterogeneity between genes and to allow for comparisons. In haptophytes all of these genes are plastid-encoded. When comparing the tree with one rate parameter and a tree where the branch leading to the haptophyte has a different rate parameter the difference in likelihood is low (Fig. IV.3, Table IV.1). In addition, with the exception of *rpl33*, *psaC* and *clpC* major differences in the rate of evolution between genes were not present, at least when compared with the dinoflagellate rates.

In dinoflagellate minicircle and plastid-transferred genes the likelihood ratio tests also show a striking increase in the evolutionary rate and large differences in likelihood (Fig.IV.3). In dinoflagellates the addition of another rate parameter has a large effect on the difference in likelihood with all minicircle and plastid-transferred genes (Table IV.1). Again, as with the relative rate tests, it is difficult to assess saturation especially because with *psbB*, *atpA* and *petD* the rate parameter reached the maximum value allowed by the program. Generally, the minicircle genes have the highest rate parameter and difference in likelihood, with plastid-transferred genes having a lower rate parameter and smaller difference in likelihood. However, *atpH*, *rpl33*, *atpI*, and *psaC* had a rate that was similar to the slower minicircle genes.

The nuclear-transferred genes *psbO* and *GAPDH* have the lowest rate parameters and likelihood differences. The values for dinoflagellate and haptophyte likelihood ratio tests using these nuclear-transferred genes were comparable.



**Table IV.1 Results of likelihood ratio tests**

The Rhino program was used without third codon position. This table is sorted by the relative rate of the dinoflagellate.

Gene name	rel. rate <sup>a</sup> haptophyte	rel. rate <sup>a</sup> dinoflagellate	$\Delta\ln^b$ haptophyte	$\Delta\ln^b$ dinoflagellate	Category <sup>c</sup>
psbB	1.01	1000.00	0.00	-389.63	minicircle
atpA	1.03	1000.00	-0.01	-269.25	minicircle
petD	1.24	1000.00	-0.37	-67.30	minicircle
petB	1.54	838.06	-0.88	-61.23	minicircle
psbE	1.21	86.06	-0.03	-36.01	minicircle
psaA	1.44	79.28	-5.47	-606.19	minicircle
atpB	1.04	51.71	-0.05	-266.36	minicircle
psbA	0.84	27.41	-0.28	-38.24	minicircle
psbD	1.31	19.81	-0.41	-61.51	minicircle
psbC	1.49	11.59	-2.04	-165.14	minicircle
atpH	1.60	244.91	-1.03	-60.99	plastid-transferred
rpl33	205.07	183.67	-4.64	-22.97	plastid-transferred
atpI	1.36	17.87	-1.26	-60.47	plastid-transferred
psaC	6.05	11.69	-3.98	-23.41	plastid-transferred
psbL	0.30	6.40	-1.65	-12.07	plastid-transferred
psaF	0.99	5.69	-0.73	-41.65	plastid-transferred
rpl3	0.86	5.69	-0.46	-4.00	plastid-transferred
rps2	0.56	5.02	-5.71	-5.71	plastid-transferred
psaD	0.69	3.88	-0.92	-21.83	plastid-transferred
psbK	0.26	3.03	-2.08	-4.00	plastid-transferred
clpC	0.15	2.69	-21.02	-11.39	plastid-transferred
chlI	1.02	2.34	0.00	-11.72	plastid-transferred
atpG	0.98	2.33	0.00	-10.48	plastid-transferred
psbO	0.88	1.01	-0.47	0.00	nuclear-transferred
GAPDH	1.41	1.70	-3.80	-1.76	nucleus-transferred

a - relative rate as calculated by the Rhino program.

b - difference in likelihood between a tree with one rate and a tree with a separate rate for the dinoflagellate or haptophyte respectively.

c - category of gene minicircles are presumed to be plastid-encoded, plastid-transferred are dinoflagellate specific plastid to nucleus gene transfers, and nucleus-transferred are plastid genes presumably encoded in the nucleus of the previous host.

#### IV.4. Discussion

The dinoflagellate chloroplast is unusual because unlike known chloroplasts only a few genes appear to have been retained in the chloroplast. The source of the dinoflagellate plastid is likely to have had a chloroplast genome with about 100 – 200 protein-coding genes. This hypothesis is supported because all of the potential sources of the dinoflagellate plastid, the red algae (Takishita and Uchida 1999; Zhang et al. 1999), heterokonts (Durnford et al. 1999), or haptophytes (Yoon et al. 2002) have chloroplast genomes with similar gene content (Kowallik et al. 1995; Reith and Munholland 1995; Grzebyk et al. 2003; Sanchez-Puerta et al. 2004). If this is correct, then many genes must have been transferred from the dinoflagellate plastid to the nucleus after (or as) the plastid was acquired.

Comparing the evolutionary rate of minicircle genes retained in the plastid with the rate of plastid-transferred genes shows that minicircle genes are evolving more quickly than genes that have been transferred (Figs IV.2, 3 Table IV.1). The likelihood ratio test depends on a tree to generate a rate of evolution and this test clearly separates the minicircle genes from the majority of plastid-transferred genes. However, both the likelihood ratio test and relative rate test indicate that some of these plastid-transferred genes including *rpl33*, *atpI*, *atpH*, and *psaC* have a rate of evolution approaching minicircle genes. In haptophytes where all these genes are retained on the plastid genome only *rpl33* and *clpC* show a pattern of accelerated evolution using both tests and the remaining genes show a relatively consistent rate of evolution.

Hypotheses that could explain the high rate of dinoflagellate chloroplast genes include relaxed selection, small effective population size, and accelerated substitution rates either in the nucleus or chloroplast or both. The hypothesis that fits the data more exactly should account for a high evolutionary rate found only in genes that are likely to have been plastid-encoded at the time of endosymbiosis and not to all plastid-associated genes. Therefore, any hypothesis invoking relaxed selection or small effective population size has to apply only to genes that are or were plastid-encoded. Although these hypotheses cannot be ruled out, a simpler explanation would be an organelle specific accelerated substitution rate, perhaps because of a low fidelity replication mechanism or increased mutation rate within the plastid. Moreover, the observed heterogeneity in rates between plastid-transferred genes is large. Although gene-specific rates of evolution could account for some differences between genes these same genes are relatively homogeneous within haptophytes and heterokonts.

One hypothesis that could explain the diversity of rates between plastid-transferred genes would be that some genes were retained in the plastid for a longer time compared to other genes and were therefore subject to the accelerated rate of the plastid. Indeed, *atpH* and *psaC* are genes otherwise only found in chloroplast genomes suggesting relatively strong selective pressure to retain these genes in the chloroplast (Race et al. 1999). Genes that were transferred earlier could have accumulated fewer mutations in the nucleus than genes that were transferred to the nucleus later after they had accumulated mutations in the chloroplast genome. It seems likely that the nucleus is a more conservative environment for these plastid genes.

If this model is correct then the evolutionary rate of plastid-associated genes that were transferred from the previous host nucleus to the dinoflagellate nucleus directly would not be affected by the high rate of the chloroplast genome. Indeed, the two putative nuclear-transferred genes used in this study, *psbO* and GAPDH, have a relative rate that is very similar to the relative rate of haptophytes and heterokonts. It is possible that *psbO* was still retained in the chloroplast genome at the time of endosymbiosis, but this seems unlikely given that *psbO* is a nuclear-encoded gene family in known photosynthetic eukaryotes (Ishida and Green 2002). For the GAPDH gene, chloroplast to nucleus transfer is unlikely because in the dinoflagellates (Fagan et al. 1998), haptophytes (Harper and Keeling 2003), heterokonts (Fast et al. 2001) and cryptophytes (Liaud et al. 1997) the cyanobacterial version of this gene is thought to have been lost and a substitute cytosolic version has been targeted to the plastid. The GAPDH evolutionary history suggests direct transfer from the nucleus of the previous plastid host to the nucleus of the new host (Fast et al. 2001; Harper and Keeling 2003).

These data suggest that a high evolutionary rate is present in the chloroplast of dinoflagellates, but that the genes acquired directly from the previous host nucleus were protected from this high rate of evolution. The extreme rate of evolution in dinoflagellate chloroplasts is presumably correlated with the observed massive decrease in plastid-encoded genes in dinoflagellates. The apparent selective advantage of retaining a specific gene in the chloroplast genome must be balanced with the potential for accumulating deleterious mutations in the chloroplast genome. In dinoflagellates this balance appears to be tilted towards gene transfer, although a few genes are still retained in the plastid. It is not yet clear whether the high rate is caused by or a consequence of

the unique minicircular DNA in the chloroplast; nor is it clear if the rate of mutation in peridinin-containing plastids is a cause for their frequent replacement by other plastid types.

#### **IV.5. Materials and methods**

Trees were constructed using PAUP\*4b10 as described in chapter III. Briefly, alignments were constructed using the known chloroplast genomes of *Arabidopsis thaliana* (Sato et al. 1999), *Nephroselmis olivacea* (Turmel et al. 1999), *Chaetosphaeridium globosum* (Turmel et al. 2002), *Mesostigma viride* (Lemieux et al. 2000), representing the green primary lineage, and *Porphyra* (Reith and Munholland 1995), *Cyanidium caldarium* (Glockner et al. 2000), *Cyanidioschyzon merolae* (Ohta et al. 2003), representing the red primary lineage, *Guillardia theta* (Douglas and Penny 1999), *Odontella sinensis* (Kowallik et al. 1995), and the unpublished data from *Emiliana huxleyi* represent secondary plastids derived from red algae. *Synechocystis* (Kaneko et al. 2001) and *Nostoc* (Kaneko et al. 2001) were used as the cyanobacterial outgroup as well as the glaucophyte *Cyanophora paradoxa* (Stirewalt et al. 1995). Data from *Amphidinium carterae* were downloaded from the dinoflagellate EST web page (<http://oxrid.umd.edu>) for plastid-transferred genes and the minicircle data from *Amphidinium operculatum* was used. The EST data, as well as Hiller (Hiller 2001), suggest that *Amphidinium operculatum* and *Amphidinium carterae* CCMP 1314 have the same minicircle gene sequences.

The genes that were directly transferred from the nucleus to the nucleus, *psbO* and GAPDH were downloaded from ncbi using the sequences published by Ishida and Green (2002) for *psbO* and Harper and Keeling (Harper and Keeling 2003) for GAPDH.

Although the GAPDH used here is the byproduct of a cytosolic gene substitution for a cyanobacterial gene, in these taxa, it is plastid-targeted and is likely to have been subject to nucleus to nucleus gene transfer. In these genes haptophyte sequences from *Isochrysis* were used for comparison with *Amphidinium* and *Heterosigma* was substituted for *Odontella* for the heterokont.

Alignments were guided by clustalw amino acid alignments and were manually adjusted as nucleotide alignments using MacClade 4.0 (Maddison and Maddison 2000). For all genes and analyses the third codon position was removed. Distance matrices were assembled by estimating GTR + I +  $\Gamma$ 4 parameters from the best log det distance tree using minimum evolution. The relative rates were calculated by dividing the dinoflagellate distances by the haptophyte distances and by dividing the heterokont distances by the haptophyte distances. The Rhino program (Rambaut, A. 2002; Rhino v1.1. available at <http://evolve.zoo.ox.ac.uk/>) was used to compare the likelihood score of trees with a single rate parameter for all taxa to a tree with two rate parameters, one for the group of interest and another for the rest of the taxa. The most likely tree was calculated with a heuristic search using maximum likelihood with the GTR model of evolution (parameters were estimated from the log det distance tree). For calculation of the dinoflagellate rates using minicircle genes and formerly plastid-encoded genes, the dinoflagellate branch (only one dinoflagellate was present) was allowed to have a different rate from the rest of the tree and the likelihood score and relative rate was calculated using Rhino. For calculation of the haptophyte rate and likelihood difference dinoflagellates were excluded and the most likely tree was found using PAUP as above, then the score of the tree with a single rate parameter and the tree with an additional rate

parameter for the haptophyte were compared using rhino. For the calculation of nuclear to nuclear transfer gene rates trees were constructed in the same way, but when more than one dinoflagellate or haptophyte was present all of the dinoflagellates or haptophytes were given the new rate parameter.

## **Chapter V Chromophyte plastid phylogeny: A plastid-encoded gene suggests that the peridinin dinoflagellate plastid originated from haptophytes**

### **V.1. Abstract**

Chloroplast characters define the chromophytes or the chlorophyll *c* containing algae, yet the phylogeny of the hosts and endosymbionts (chloroplasts) is not well understood. The plastids of chromophyte algae were gained either once, implying host monophyly, or multiple times by different host lineages. We have used the chloroplast gene *psbB* to construct a phylogenetic tree including all the chromophyte lineages. The maximum likelihood tree places the dinoflagellates within the haptophyte algae with moderate nonparametric bootstrap support. The dinoflagellate plastid genes are very divergent and this can lead to long branch artifacts. Parametric bootstrapping was used to test the ability of likelihood to compensate for the asymmetric shape of the tree. The approximately unbiased test was used to test alternate tree topologies, and only trees placing dinoflagellates sister to or within haptophytes were accepted. The chloroplast gene tree implies that the dinoflagellate plastid was acquired from a haptophyte or *visa versa*. The *psbB* tree is congruent with several hypotheses of chloroplast transfer within the chromophytes.

### **V.2.1. Introduction**

The term chromophyte refers to algae pigmented with chlorophyll *c*. Taken together, these organisms constitute a large and ecologically important assemblage, including the heterokonts (i.e., stramenopiles), haptophytes, cryptophytes and dinoflagellates. These algae are responsible for substantial primary productivity in our oceans, and their blooms can be ecologically and economically devastating events.

The initial concept of the chromophytes as a taxonomic group was based on their common pigmentation, a plastid character (Christensen 1962; Christensen 1989). The very fact that chromophytes share a distinctive pigmentation -- in most cases chlorophyll *c* and xanthophylls are the light-harvesting pigments -- suggests a common plastid origin. All chromophytes have secondary plastids derived from red algae, but it remains uncertain how many independent acquisitions of plastids have occurred in these organisms (Daugbjerg and Andersen 1997; Delwiche and Palmer 1997; Martin et al. 1998; Douglas and Penny 1999; Durnford et al. 1999; Oliveira and Bhattacharya 2000; Yoon et al. 2002). Since plastids are endosymbiotic organelles plastid and nuclear phylogeny need not be congruent. This requires a revision of the chromophyte concept: are chromophytes a monophyletic group of plastids in a diverse array of hosts?

Either the plastids were acquired separately in the host lineages (requiring up to four plastid gains), or plastids were acquired once in the ancestral host lineage (implying host monophyly). There are a number of possible intermediates between these two extremes and discriminating between these hypotheses requires strong host and plastid phylogenies.

To understand better the relationships among chromophyte and red algal plastids, we sequenced the gene *psbB*, which encodes the ca. 508 amino acid chlorophyll *a* binding light-harvesting protein (CP47 or CpA-1) of the photosystem II complex (Rhee 2001) from several dinoflagellates, haptophytes and heterokonts and we used these for phylogenetic analysis. This protein has six transmembrane regions, and binds roughly 20 chlorophyll *a* molecules and  $\beta$ -carotene (Bricker 1990). The gene *psbB* is encoded in the plastid genome of all chromophytes including dinoflagellates, and its rate of sequence evolution makes it promising for molecular phylogenetic studies.

Plastid-encoded genes from dinoflagellates are encoded by small plasmids, or minicircles, and seem to have an extremely high rate of sequence evolution (Takishita and Uchida 1999; Zhang et al. 1999; Barbrook and Howe 2000). Maximum likelihood (ML) is a method that uses an explicit model of sequence evolution and is known to perform more consistently under conditions of highly unequal rates of sequence evolution (Felsenstein 1978; Swofford et al. 1996). In our study we use parametric bootstrapping (Huelsenbeck et al. 1996) to test the ability of likelihood methods to recover phylogenetic trees under conditions comparable to those seen with the dinoflagellate data. The approximately unbiased test was used to assess whether differences in the tree topology produced trees with significantly different likelihood. These methods explicitly question the underlying assumptions of the analysis, but instead of undermining the result they place accurate limits on the ability of our data to discriminate between hypotheses.

### V.3.3. Results

The *psbB* alignment is unambiguous and, when dinoflagellates are excluded, continuous from the start to stop codon. There is some length variation in the dinoflagellates, with *Amphidinium operculatum* encoding a 506 aa protein, *A. sanguinea* 502 aa, and *Gyrodinium impudicum* roughly 490 aa (cf. *Nicotiana* 508aa). The *G. impudicum* PCR product was unusual, with a 29 bp deletion that would result in a frame-shift mutation. Despite this deletion, the sequence is readily aligned with the other sequences, but the aligned sequence does include a single in-frame stop codon. This was the only amplification product obtained from *G. impudicum*.

The dinoflagellate sequences are somewhat unusual among *psbB* sequences. Of the fourteen conserved histidines in *psbB* which are thought to be chlorophyll binding 12 are present in all of the taxa examined, and two have been substituted in dinoflagellates: histidine 201 is arginine in some dinoflagellates and histidine 343 is glutamine in *A. operculatum* (numbering after Bricker 1990).

The total usable length of the nucleotide alignment was 1387 bases. Most plastid DNA sequences are AT rich, and this alignment is no exception; the average AT content for the putative plastid-encoded sequences in the alignment was 59.3%. The dinoflagellate sequences have a similar base composition on average (59.5%), but that of *A. operculatum* is unusually low (50.1%).

The parsimony tree (not shown) grouped the dinoflagellates and the cyanobacteria with strong bootstrap support. The branching order of the other chromophyte lineages is similar to the likelihood tree. The distance trees using ML distance matrices group the dinoflagellates with the haptophytes, although the relationship of this clade to others was

sensitive to model parameters, particularly to the fraction of invariant sites. Distance comparisons between dinoflagellates and the other taxa are nearly all above one, and some within dinoflagellate distances approach two. However, after the parameters were optimized, the distance tree was generally topologically congruent with the likelihood tree. The GTR + I +  $\Gamma$  (4 categories) model of sequence evolution was found to be the optimal by the likelihood ratio test. The likelihood of this tree did not improve with the addition of more than four rate categories. A comparison of raw and maximum likelihood distances using the optimum model (GTR + I +  $\Gamma$ ) is shown in figure V.1.

The maximum likelihood tree found using the GTR + I +  $\Gamma$  model is shown in Fig. V.2, with nonparametric bootstrap values shown above the branches. Although the likelihood ratio test indicates that the GTR + I +  $\Gamma$  model is significantly better than GTR +  $\Gamma$  model ( $\Delta\ln = 12$ ), the GTR +  $\Gamma$  tree is topologically identical to the tree found using the optimal model, and the two trees differ only in mean branch length differences of 0.4%, with a standard deviation of 0.3% (the maximum branch length discrepancy observed was less than two percent).

Some key clades show robust nonparametric bootstrap support: the cyanobacterial outgroup is monophyletic (100%), with moderately strong (86%) support for placement of the glaucophyte *Cyanophora paradoxa* as the sister to that clade; the green algae (including *Euglena*; 100%); and the red algae (95%), which are embedded in a clade with the chromophytes (94%). Within the red algal/chromophyte clade, the heterokonts *Odontella* and *Pelagomonas* form a moderately supported (70%) monophyletic group, while the third heterokont, *Chattonella* (a raphidophyte), is separated from this group by a very short branch and unsupported (<50%) branch that places it in a clade with

*Guillardia* (a cryptomonad) and the red algae. Sister to this group is a clade composed of haptophytes with a monophyletic group of dinoflagellates embedded. Monophyly of the dinoflagellates is strongly supported, but the branching order within the haptophytes finds only weak to moderate support (68% - 78%).

In likelihood analyses of the parametric bootstrap data most features of the tree were recovered in all replicates, with the exception of four branches, as shown by grey branches and circled values on figure V.2. The node separating the *Gyrodinium* and *Akashiwo* clade from *Gymnodinium simplex* was recovered in 93% of the replicates. *Odontella* and *Pelagomonas* were placed within the red algal clade in 4% of the analyses. The two other nodes were recovered in 98% and 99% of the trees. All other features of the tree, including the clade composed of dinoflagellates and haptophytes, were recovered in 100% of parametric bootstrap replicates.

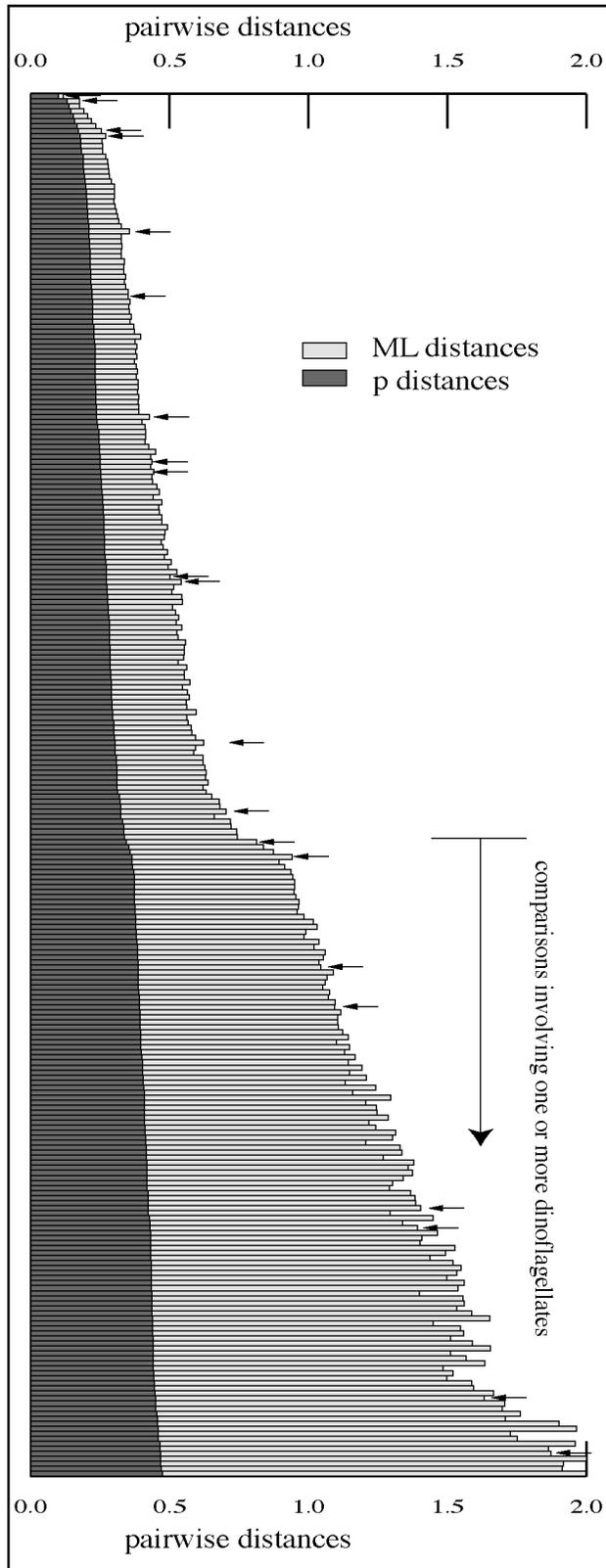
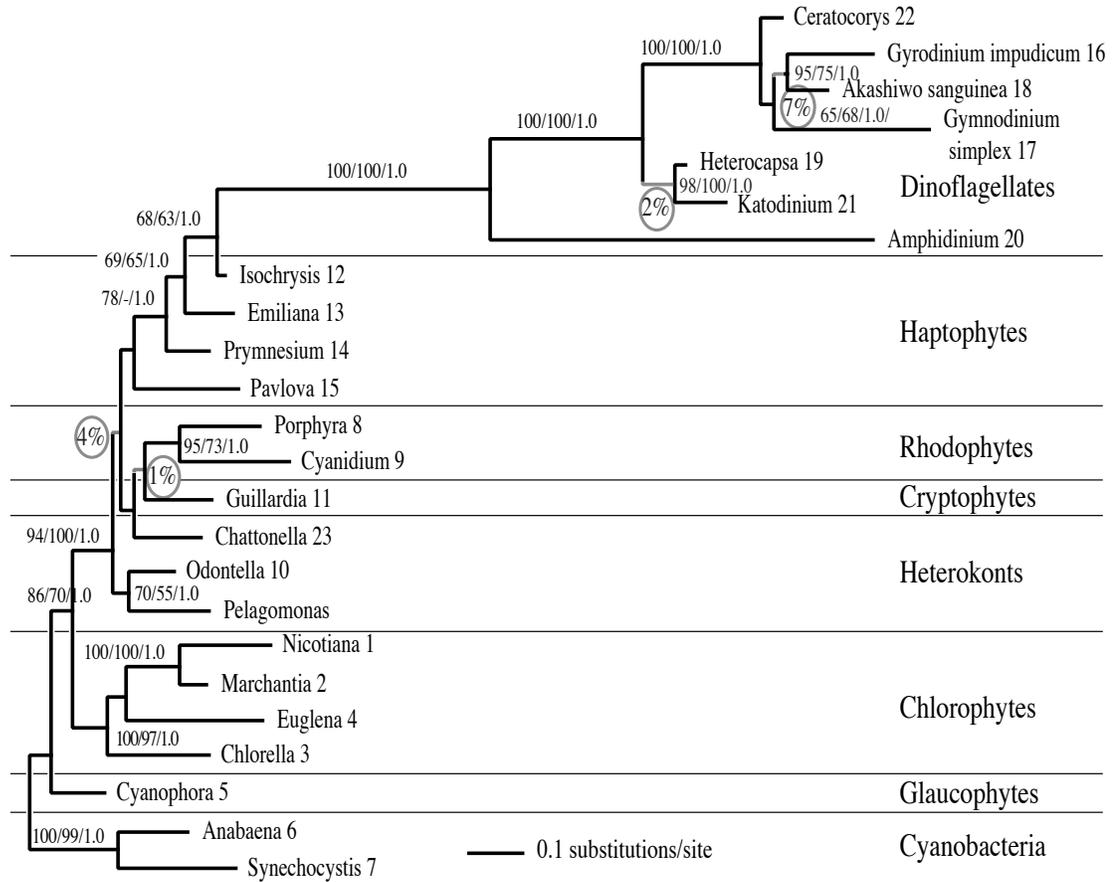


Figure V.1: P-distances and maximum likelihood (GTR + I +  $\Gamma$ ) distances compared for all pairwise combinations of taxa, and ranked by P-distance. P-distances are the number of substitutions divided by the length of the sequence, and maximum likelihood distances are corrected for superimposed substitutions. Those distances that correspond to comparisons of two dinoflagellates are marked by an arrow, and those that involve at least one dinoflagellate – which constitute the entire upper end of the range – are indicated. Most distances involving dinoflagellates are above one substitution per site.

Figure V.2: Maximum likelihood tree for psbB, based on the GTR + I +  $\Gamma$  model. Nonparametric bootstrap proportions above 55% are shown above the corresponding branch. The first proportion is when all data is used, the second when the third position is excluded (where the trees are compatible) and the third is the posterior probability of the MrBayes tree when a single GTR + I + G model is used. The four branches that had parametric bootstrap proportions of less than 100% are shown in grey, with the corresponding parametric bootstrap proportion circled. Numbers after the taxon names refer to the trees presented in table V.3.



When parametric bootstrap analysis was performed using the tree found by parsimony analysis, all features of this tree were found in 100% of replicates, including the placement of dinoflagellates with cyanobacteria (data not shown).

Bayesian methods produced trees that placed the dinoflagellates in the same position relative to the haptophytes with strong support with different models (i.e. covarion, multiple models for each site, third position excluded), but placed the *Odontella + Pelagomonas* clade as sister to *Prymnesium* and placed *Pavlova* and *Chattonella* in a deeper branching clade with the red algae and *Guillardia*. Bayesian posterior probabilities (Huelsenbeck and Ronquist 2001) using all positions and a single GTR + I +  $\Gamma$  (4 categories) model are shown on figure V.2. When the third codon position was excluded, two models were used for the first and second position, and the covariate parameter was on a novel tree placing the red algae at the base of a monophyletic chromophyte clade with *Guillardia* branching first, *Chattonella* branching next, and finally *Odontella* and *Pelagomonas* as a sister group to the haptophyte dinoflagellate clade was found.

Given the long branches separating some of the taxa, and the potential for third position saturation, analyses were performed using amino acids, and excluding the third codon position. Excluding the third position moved *Pavlova* out of the haptophyte clade, and embedded *Chattonella* between *Guillardia* and the red algae. This analysis also had lower bootstrap support for several key clades (values shown on Fig. V.2). However, the relationship of the other haptophytes and dinoflagellates is not altered by removing the third position.

PAML (Yang 1997) was used both for optimal tree searching and hypothesis testing using the Wagner matrix with the shape parameter set to 1.03, this parameter being optimal in an estimate using the most likely tree found by PAUP. With random taxon addition PAML (and MOLPHY) converged on trees that placed the dinoflagellates with cyanobacteria. When PAML was given the optimal ML nucleotide tree (Fig. V.1) as a starting tree it found an optimal tree placing the dinoflagellates with the haptophytes. This tree was used for hypothesis testing as item 14 (Table 3), although it is clearly not the most likely PAML tree (item 8). It is quite likely that the globally optimal PAML tree was not found. For likelihood scoring the optimal constrained trees from a PAUP search were used; tree and site likelihoods were calculated using the parameters described above.

The approximately unbiased test results are shown in Table V.3 and the constrained trees are shown in figure V.3. The confidence set of trees using a 0.10 confidence interval when PAUP was used include trees that showed different relationships of haptophytes and dinoflagellates. Accepted hypotheses (trees) include: the dinoflagellates were excluded from the terminal haptophyte clade and placed sister to *Pavlova* (item 2), haptophytes were constrained to monophyly (item 3), the most likely tree using the first and second position (item 4), the best tree placing dinoflagellates sister to *Prymnesium* (item 5). The hypothesis that constrained dinoflagellates as sister group to *Emiliana* was not accepted (item 6). Other accepted hypotheses make the chromophytes monophyletic (item 7), the heterokonts monophyletic (item 8) and the dinoflagellates, heterokonts and haptophytes monophyletic (item 9). In all accepted hypotheses dinoflagellates are embedded within haptophytes in the same way as the

optimal tree (items 1, 4, 7, 8) when this relationship is allowed to vary, but a sister relationship of the two clades or of dinoflagellates and *Prymnesium* or *Pavlova* is also accepted by this test (items 2, 3, 5). The results are remarkably similar when the third codon position was excluded and site likelihoods were used for the same fourteen trees except that the hypothesis placing the dinoflagellates with *Prymnesium* was not accepted. When the PAML package was used to score the same trees, the results are substantially different. The most likely tree still places the dinoflagellates with haptophytes, when the heterokonts are constrained to monophyly (item 8). However, alternate hypotheses placing the dinoflagellates with the heterokonts (item 10) and with the cyanobacteria were accepted (item 11), while hypotheses placing the dinoflagellates in different positions relative to the haptophytes were not (items 2,5,6,7).

**Table V.1 Results of the Approximately Unbiased test**

Item	Descriptive Constraint	Tree or Constraint <sup>f</sup>	PAUP (all NT) <sup>a</sup>		PAML (AA) <sup>b</sup>		PAUP (NT 1&2) <sup>c</sup>	
			ΔlnL <sup>e</sup>	AU <sup>f</sup>	ΔlnL <sup>e</sup>	AU <sup>f</sup>	ΔlnL <sup>e</sup>	AU <sup>f</sup>
1	Best PAUP ML tree all data	((6,7)(5(((1,2)(4)3)((10,24)(23(8,9)1)(12-15(16-22))))))	-	<b>0.742</b>	11.5	<b>0.23</b>	12.9	<b>0.423</b>
2	<i>Pavo</i> then dinos+haptos	(15(16-22)(12-14))	2.6	<b>0.575</b>	18.1	0.025	18.2	<b>0.168</b>
3	Haptophyte monophyly	(12-15)	3.3	<b>0.398</b>	17.2	<b>0.099</b>	18.8	<b>0.123</b>
4	Best PAUP ML no third	((6,7)(5(((1,2)(3,4)((11(23(15(8,9))))((10,24)(14(13(12(16-22))))))))))	5	<b>0.476</b>	6.3	<b>0.455</b>	-	<b>0.868</b>
5	<i>Pyramnesium</i> +dinos	(14,16-22)	7.2	<b>0.124</b>	16.5	0.024	19.6	0.069
6	<i>Emiliania</i> +dinos	(13,16-22)	7.5	0.066	15.9	0.02	18.7	0.015
7	Chromophyte monophyly	(10,11,23,24,12-15,16-22)	8.7	<b>0.378</b>	16	0.066	16.4	<b>0.168</b>
8	Heterokont monophyly	(23,24,10)	11.5	<b>0.189</b>	-	<b>0.831</b>	16.8	<b>0.28</b>
9	All groups monophyletic	(23,24,10)(12-15)(16-22)	14.9	<b>0.147</b>	5.6	<b>0.249</b>	23.3	<b>0.11</b>
10	Cyanos+dinos	(6,7,16-22)	16.5	0.082	20.1	<b>0.126</b>	24.1	0.088
11	Dinos+heterokonts	((23,24,10)(16-22))	20.8	0.026	5.3	<b>0.33</b>	24.9	0.059
12	Bayesian tree <sup>g</sup>	((6,7)(3(4(1,2)))5(8,9)(11(23(10,24)(14(12,13))(15(16-22))))))	72	9.00x10 <sup>-05</sup>	49.8	0.015	76	8.00x10 <sup>-05</sup>
13	Greens+dinos	(1,16-22)	74.4	1.00x10 <sup>-44</sup>	47.8	4.0x10 <sup>-04</sup>	56.1	1.00x10 <sup>-07</sup>
14	PAML tree <sup>h</sup>	((6,7)(5((4(3(1,2)))8,9)(23(10,24)(15(11(12(13,14))(16-22))))))	416	1.00x10 <sup>-49</sup>	130.6	2.0x10 <sup>-47</sup>	185.8	2.00x10 <sup>-41</sup>

<sup>a</sup> Score and approximately unbiased score of best tree given constraint, TBR search performed by PAUP using nucleotide data and all codon positions.

<sup>b</sup> Best tree (left) scored by PAML using amino-acid maximum likelihood.

<sup>c</sup> Best tree (left) scored by PAUP using nucleotide data excluding third codon positions.

<sup>d</sup> The constraint used for the PAUP search, numbering after figure V.2 trees shown in figure V.3.

<sup>e</sup> Change in ln likelihood relative to best tree in column

<sup>f</sup> Approximately unbiased P value. Values above 0.10 are shown in boldface.

<sup>g</sup> Mr Bayes tree found when the third codon position was excluded and two models were used for the first and second position with the covarion parameter on (see figure V.3).

<sup>h</sup> PAML tree found after search using tree #1 as starting tree (see figure V.3).

Figure V.3: Trees used for the approximately unbiased test. The constrained groups are in shadowed boxes, the branching order within constrained groups is not specified.

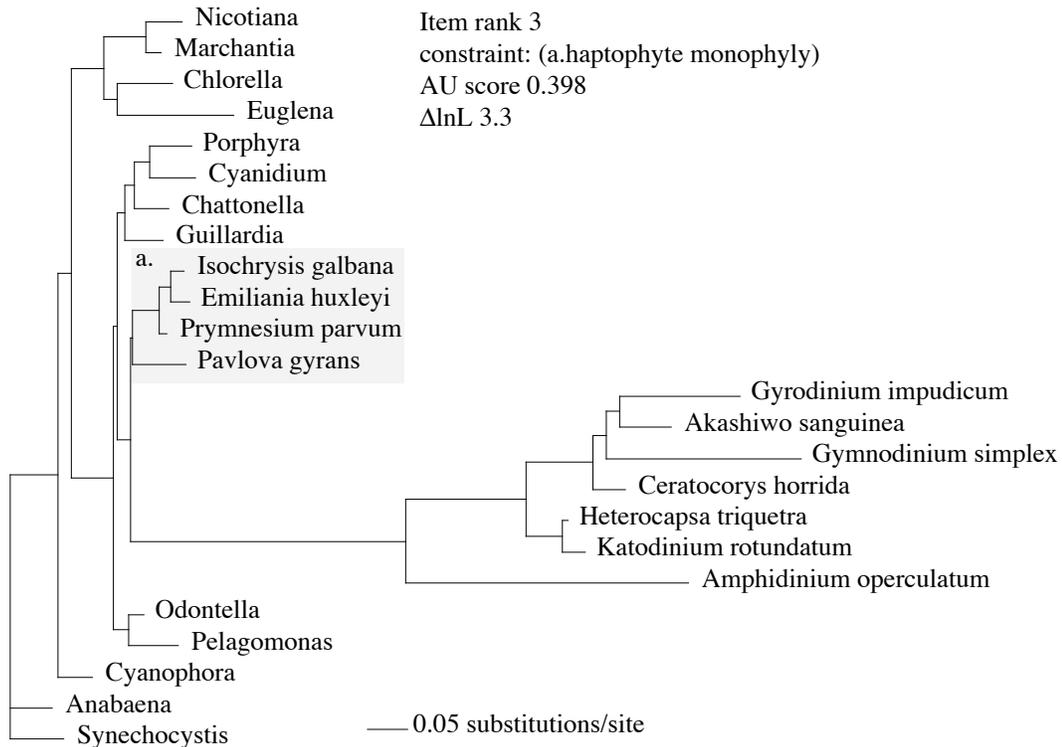
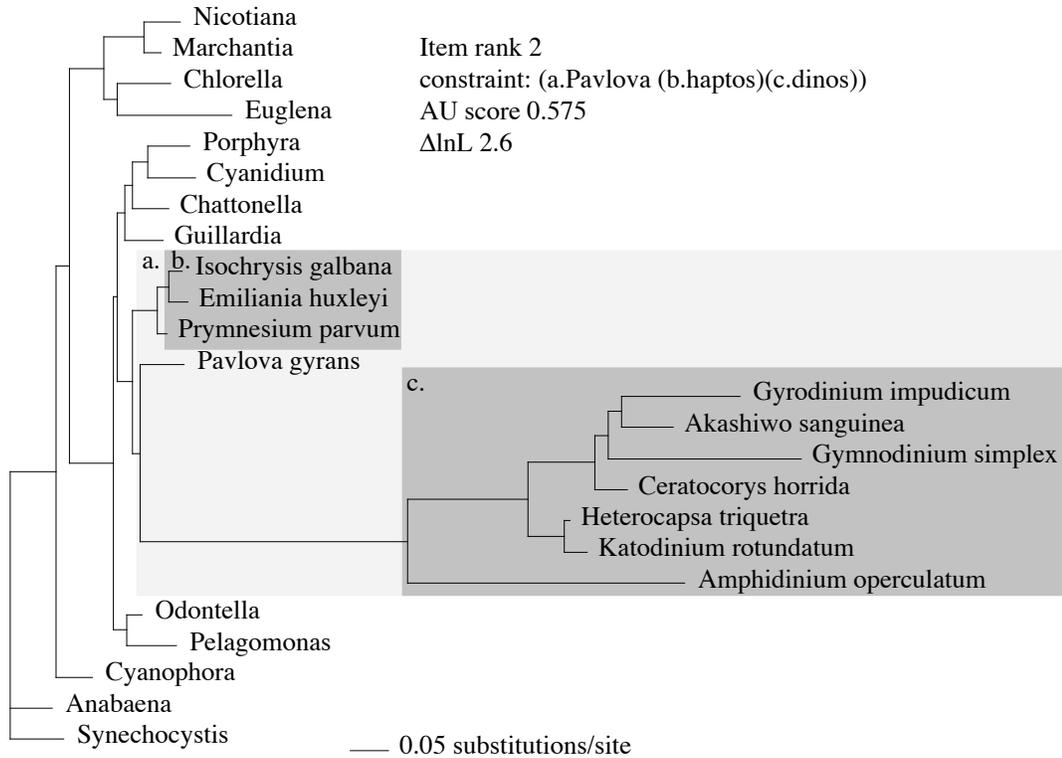


Figure V.3: (continued): Trees used for the approximately unbiased test. The constrained groups are in shadowed boxes, the branching order within constrained groups is not specified.

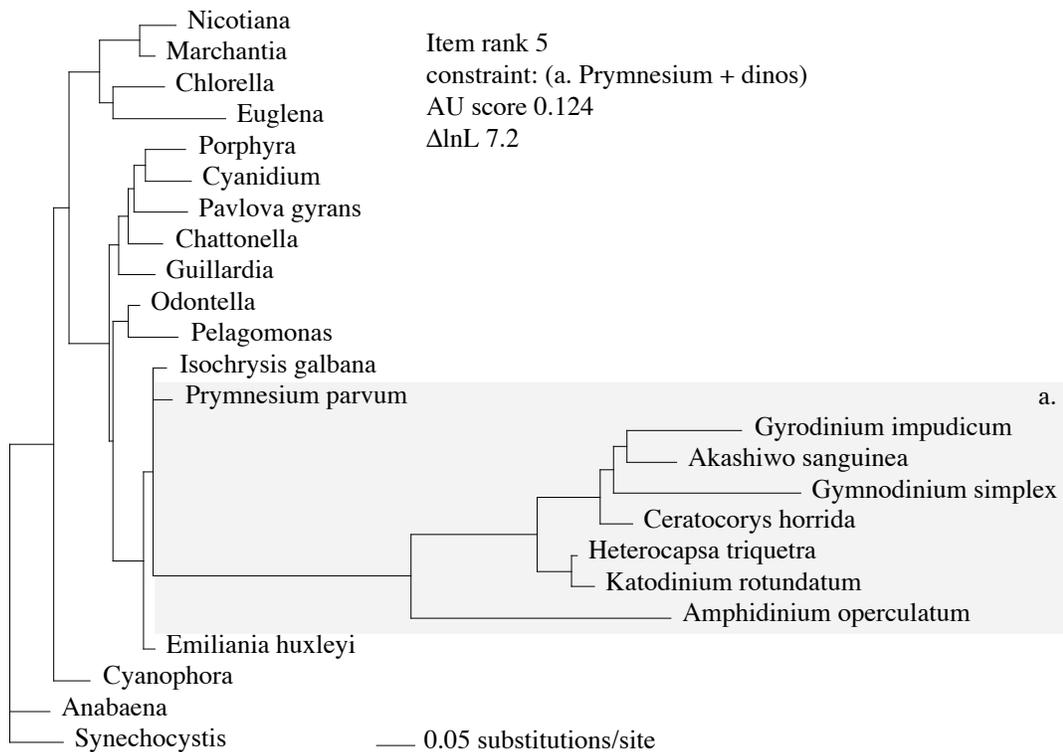
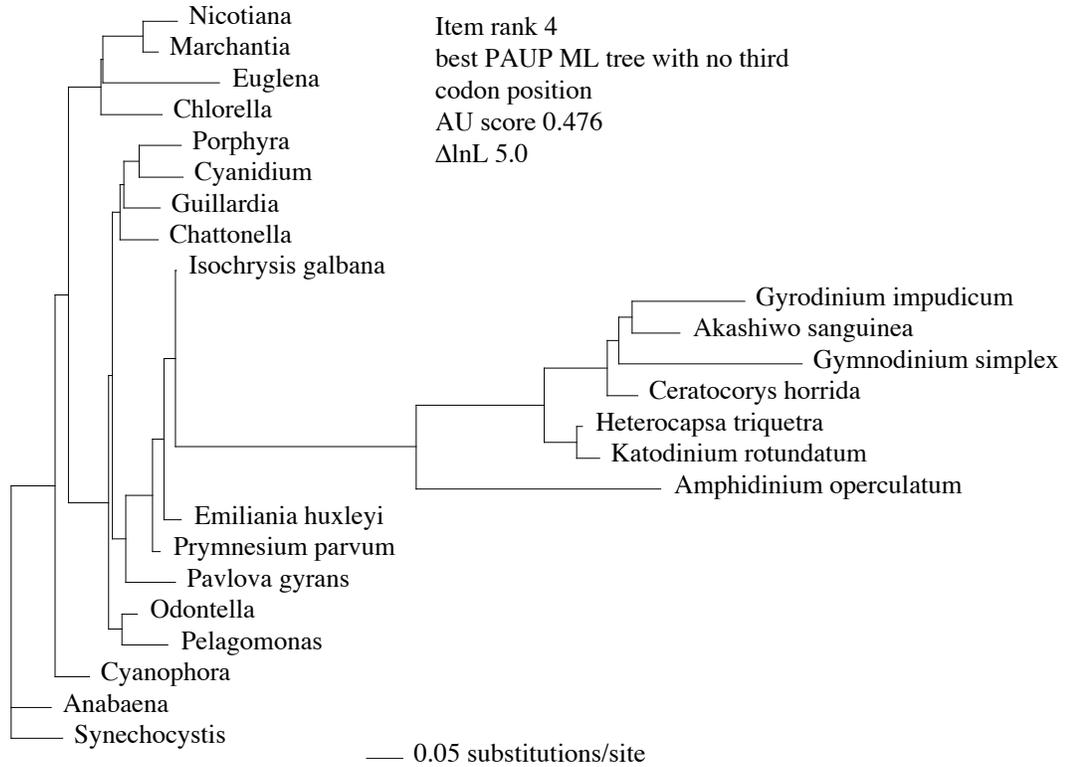


Figure V.3 (continued): Trees used for the approximately unbiased test. The constrained groups are in shadowed boxes, the branching order within constrained groups is not specified.

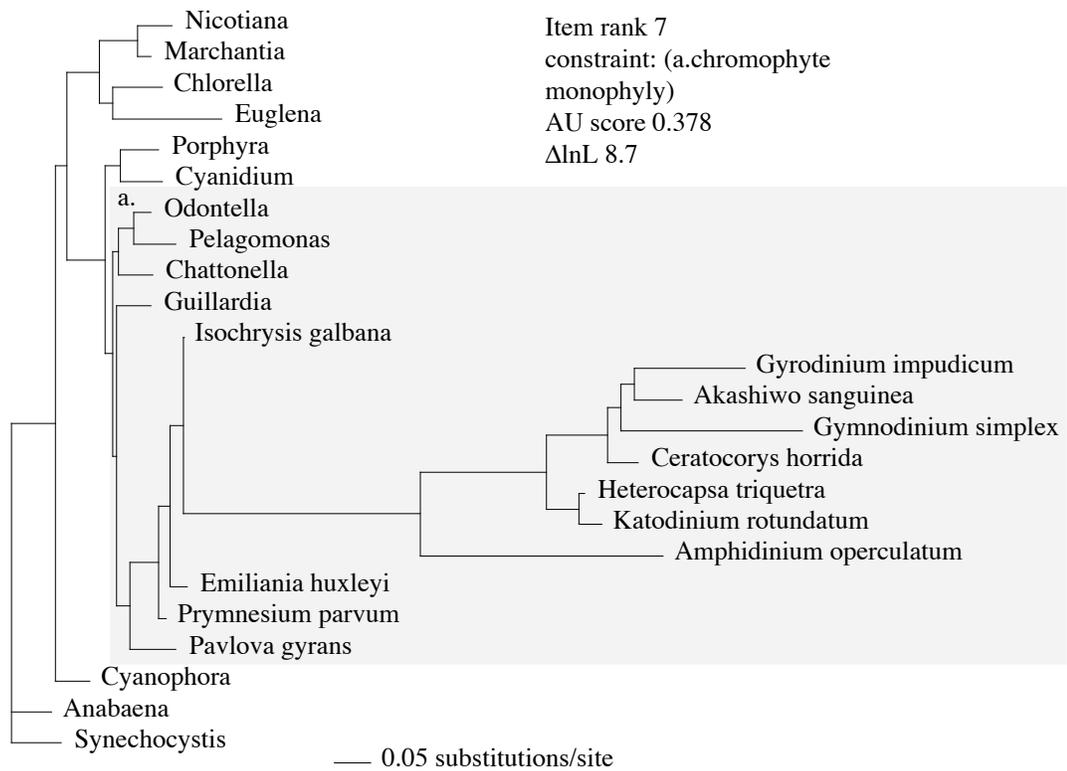
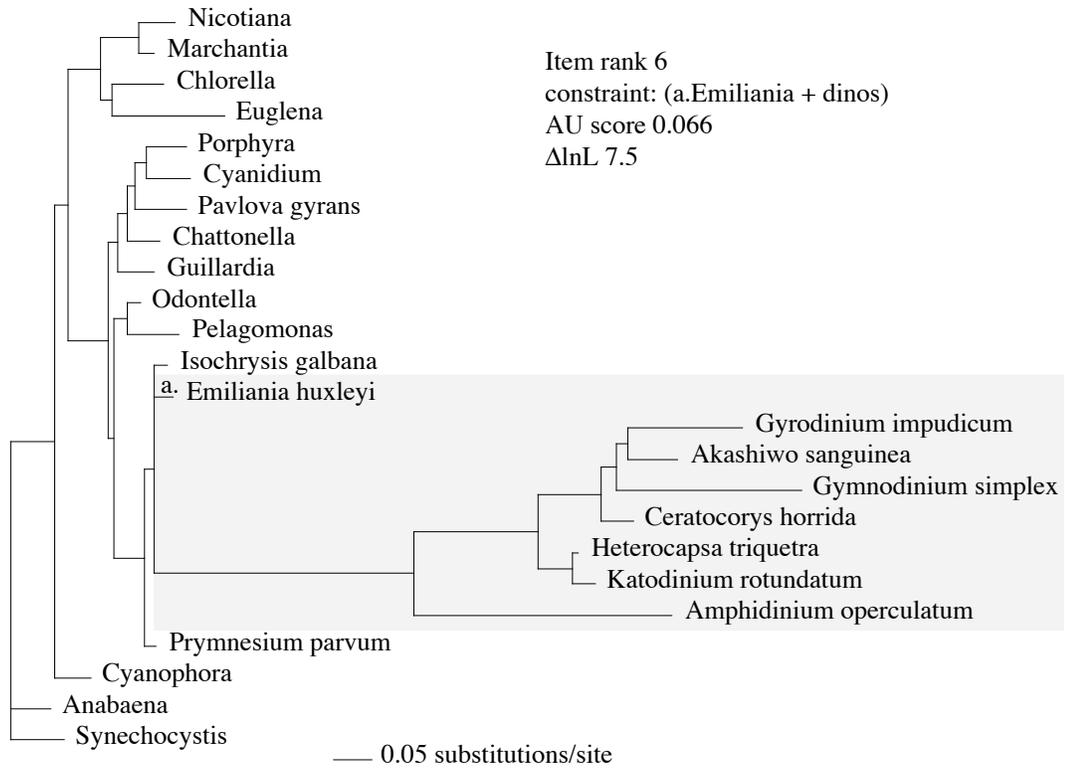


Figure V.3 (continued): Trees used for the approximately unbiased test. The constrained groups are in shadowed boxes, the branching order within constrained groups is not specified.

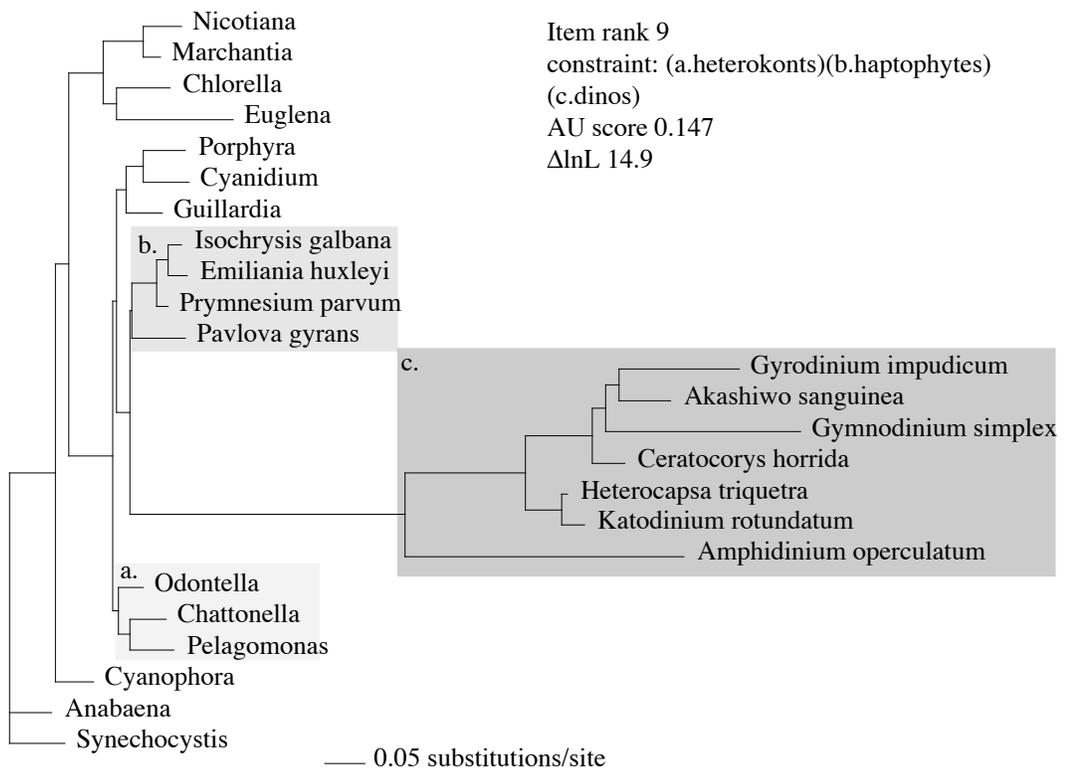
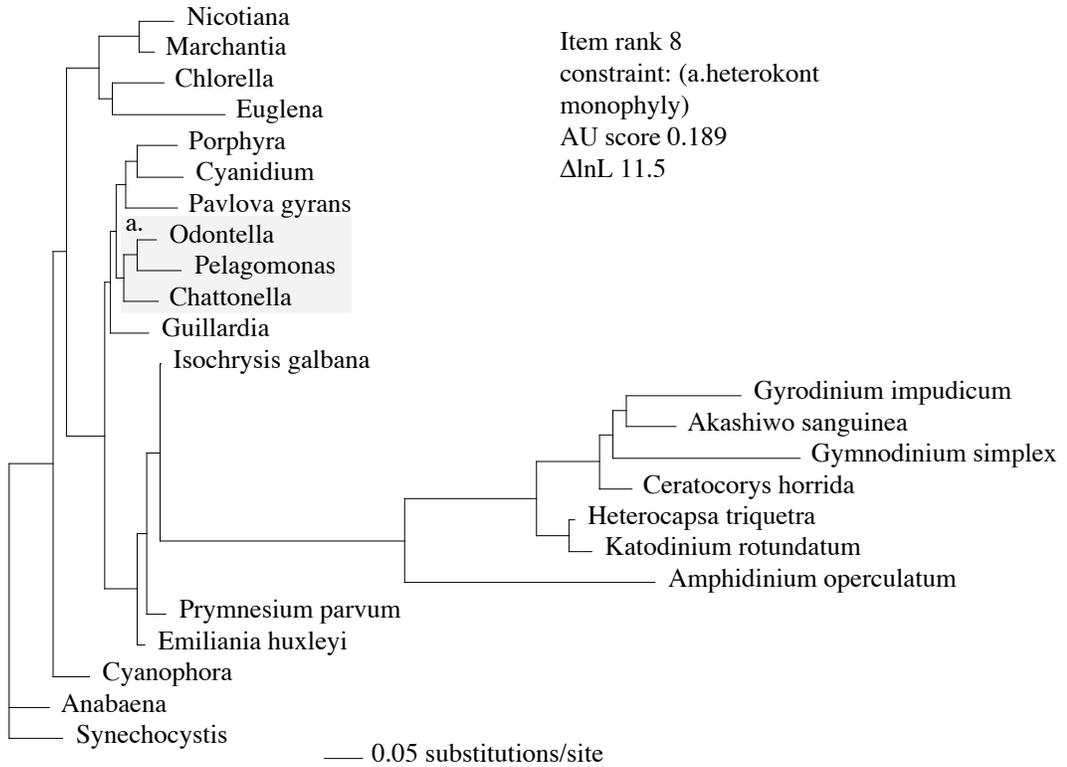


Figure V.3 (continued): Trees used for the approximately unbiased test. The constrained groups are in shadowed boxes, the branching order within constrained group is not specified.

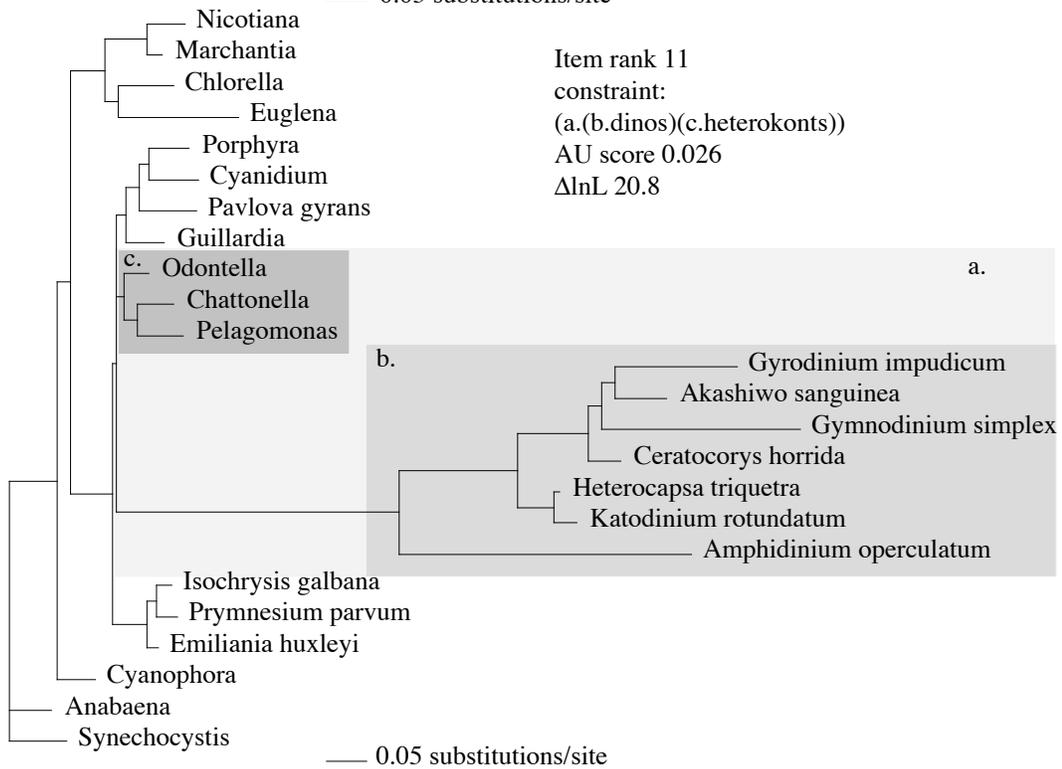
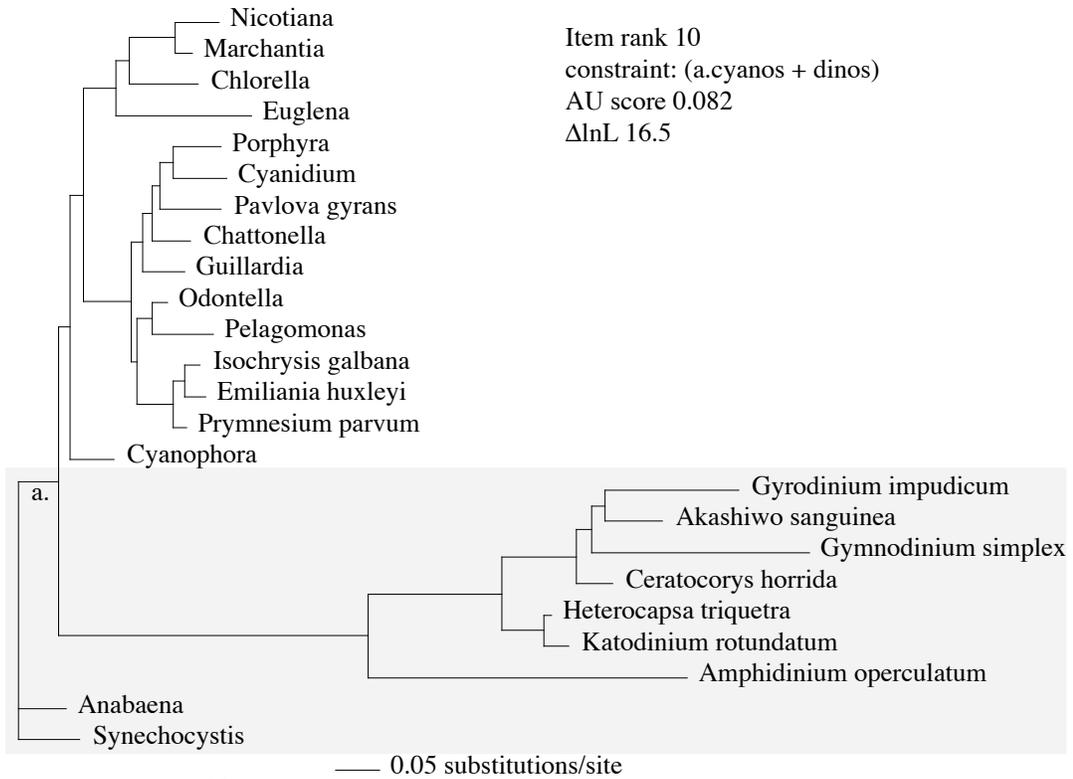


Figure V.3 (continued): Trees used for the approximately unbiased test. The constrained groups are in shadowed boxes, the branching order within constrained group is not specified.

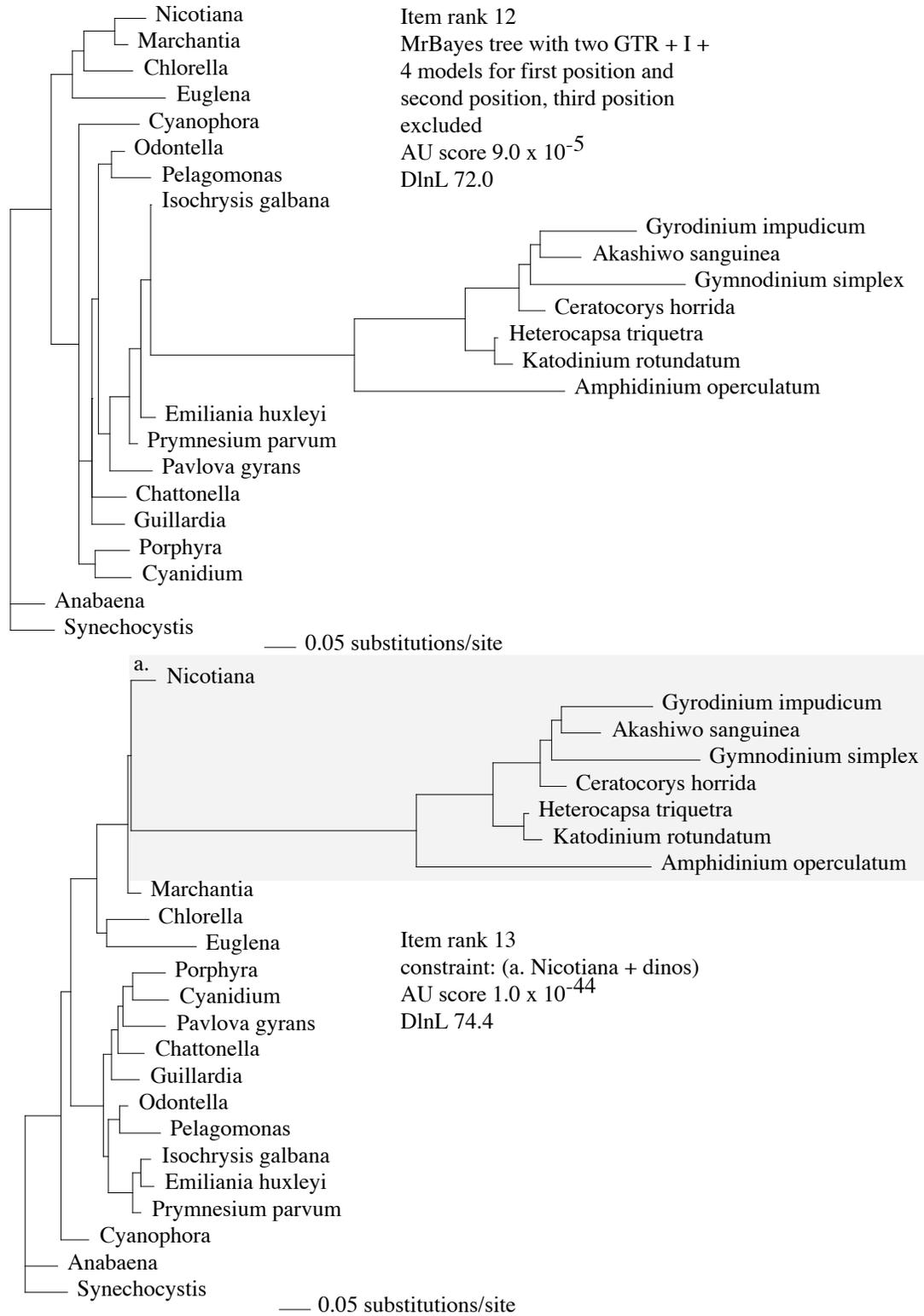
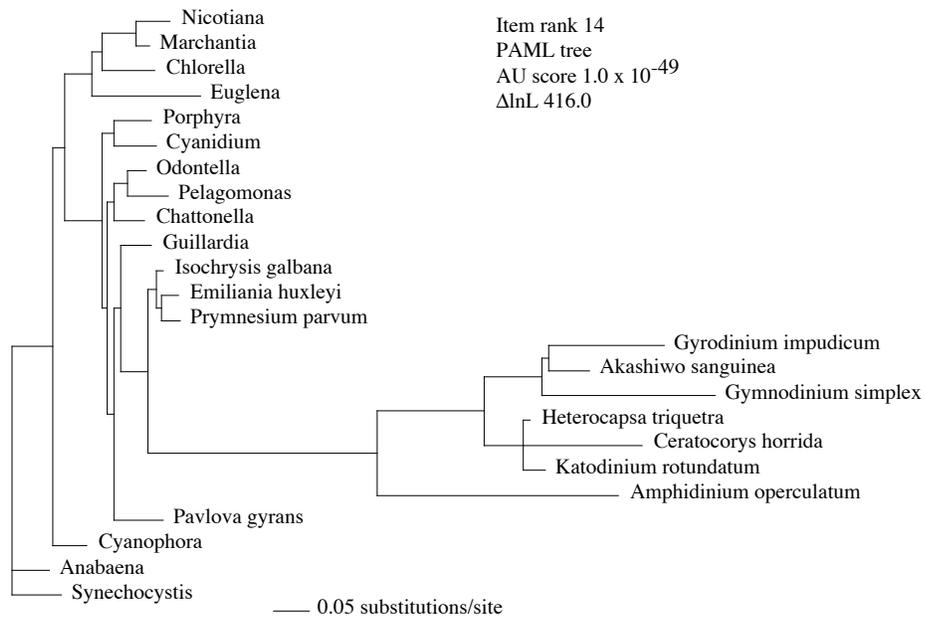


Figure V.3 (continued): Trees used for the approximately unbiased test. The constrained groups are in shadowed boxes, the branching order within constrained groups is not specified.



#### **V.4.4. Discussion**

##### *4.1. Long branch effects and parametric bootstrapping*

The ML *psbB* gene tree has two striking features: the branches leading to the dinoflagellates are extremely long, and the dinoflagellate plastid is embedded within the haptophyte plastid clade (Fig. V.2). Both findings are consistent with previous molecular phylogenies of dinoflagellate chloroplast genes, which have shown extremely high rates of evolution in analyses of several genes (Uchida et al. 1988; Takishita and Uchida 1999; Zhang et al. 1999). The crucial question regarding this phylogeny is whether any model of evolution can compensate for these extreme differences in rates. Model based methods of analysis can help compensate for the superimposed substitutions that are known to occur under conditions of high and unequal rates of sequence evolution, but no method is completely immune to these effects. Both likelihood methods and Bayesian analysis found a tree grouping dinoflagellates and haptophytes within the red algal plastid lineage (Fig. V.2). This tree has a backbone congruent with previously produced trees of plastids (Daugbjerg and Andersen 1997; Delwiche and Palmer 1997; Martin et al. 1998; Douglas and Penny 1999; Durnford et al. 1999; Takishita and Uchida 1999; Zhang et al. 1999). The branch uniting the dinoflagellates with the haptophytes has a length of 0.50 changes per site. Given the most likely tree and model of sequence evolution, half of the sites that are free to vary are inferred to have changed between these nodes. Uncorrected methods would be expected to substantially underestimate the total number of substitutions, as is dramatically shown by comparisons of maximum likelihood and raw

nucleotide (“P”) distances (Fig. V.1). This distance implies that essentially every variable site within the sequence has undergone at least some substitution.

We used the parametric bootstrap because it can provide a statistical framework to evaluate the model of sequence evolution and method of tree building. Nonparametric bootstrapping is a more widely used technique that can test the internal consistency of the data and provide a measure of whether or not there are enough data available to make a given phylogenetic inference. Nonparametric bootstrapping provides a valuable measure of information content, but is subject to a number of well-characterized limitations and biases (Hillis and Bull 1993). Most important in this context, under analytical conditions where a method is performing inconsistently traditional nonparametric bootstrapping can give artificially high values, leading to a false impression of support. Consequently, although there are several branches in the *psbB* ML tree that find strong nonparametric bootstrap support (Fig. V.2), the high rate of sequence evolution makes it difficult to interpret whether these results are caused by real phylogenetic history or problems with the model of sequence evolution.

Parametric bootstrapping uses a model of sequence evolution to generate simulated sequence alignments that correspond to the test tree. These can be used to evaluate the ability of different analytical methods to recover a tree given a particular set of analytical conditions (Huelsenbeck et al. 1996; Sanderson and Shaffer 2002). Because the same likelihood model was used for the original analysis and to simulate datasets during parametric bootstrapping, it is expected that the starting tree topology would be recovered with high frequency. In this case four branches were not always recovered during parametric bootstrapping (Fig. V.2), indicating that the model cannot always

reproduce the tree. These branches involved rearrangements among the dinoflagellates, the placement of the heterokont *Chatonella*, and the relationship between haptophytes and red algae.

By contrast, the branches that place dinoflagellate plastids among those of haptophytes were always recovered by parametric bootstrapping and found moderate nonparametric bootstrap support (Fig. V.2). It is important to note that the methods used here cannot test violations of the underlying model of sequence evolution, but these results indicate that given this model of sequence evolution the placement of dinoflagellates among haptophytes is within the scope of the analytical method. As a further test of the model the parsimony tree that placed the dinoflagellates with the cyanobacteria was used for simulation. The starting tree was always recovered, and this indicates that the likelihood model is not always misled into placing the dinoflagellates with the haptophytes, and can consistently reproduce different starting trees. In other words, if the parsimony tree were the correct phylogeny it would be expected that ML would have found it.

While most phylogenetic studies present a single most probable tree, it is perhaps more interesting and informative to compare a suite of trees, and to find a set of trees that are approximately equally likely given the data. Such an analysis measures the information content of the data without suggesting false resolution. Several tree comparison methods are available, and the difference in likelihood of the optimal tree and other trees can be compared, but the problem then becomes one of discriminating the statistical significance of this difference. The approximately unbiased test evaluates differences in a set of trees so that different trees (hypotheses) can be compared, and has

some notable advantages over the more widely used Kishino-Hasegawa test (Shimodaira 2000; Shimodaira and Hasegawa 2001; Lang et al. 2002; Douady et al. 2003). If the trees are not significantly different from each other then instead of considering one tree the ‘true’ tree, a set of trees have to be considered. This allows for both confidence tests and measures the relative power of the data and model. In this case the 90% confidence set of trees that are not significantly different from each other all placed the dinoflagellates within or sister to the haptophyte clade in different combinations (i.e. items 1-5, 7-9 in Table V.3). The confidence set of trees does not allow us to distinguish the exact position of the dinoflagellates with respect to the haptophytes. Two of the hypotheses that were rejected placed the dinoflagellates with the cyanobacteria (item 10) and with the heterokonts (item 11) respectively. When the potentially saturated third codon position is excluded (Fig. V.1), the optimal tree changes but only the tree placing dinoflagellates with *Prymnesium* is excluded from the confidence set. In this case the results correspond well to a heuristic idea: it is biologically implausible that the dinoflagellate plastid is directly derived from the cyanobacteria, and this implausible hypothesis is not accepted by this test.

The difficulty of this dataset is clear when protein methods are employed, PAML was unable to find the optimum PAML tree both when using random addition and when given a starting tree. This is not surprising considering that PAML is not optimized for tree searching (Yang 1997), but is used here to score a set of alternative trees. Given this data and the Wagner matrix of amino acid substitution (a model that is potentially suboptimal for this dataset), the approximately unbiased test cannot exclude a

relationship of dinoflagellates to either heterokonts or cyanobacteria (Table V.3), but the best scoring tree is one placing the dinoflagellates in the same position as figure V.2.

Both the parametric bootstrap and the approximately unbiased test reveal potential problems with the model of sequence evolution, and the power of the data to discriminate between hypotheses. However, these results actually add strength to the assertion that there is a relationship between dinoflagellate and haptophyte plastids, since both tests revealed that this relationship was accepted.

#### *4.2. Peridinin dinoflagellate plastids and haptophyte plastids*

A close relationship between some dinoflagellate and haptophyte plastids is suggested by a group of anomalously pigmented dinoflagellates. These dinoflagellates have the distinctive xanthophyll 19' hexanoyloxyfucoxanthin (19'), which is a characteristic pigment of haptophytes. Ribosomal data from the plastids of these dinoflagellates suggests that the 19' plastids were derived from haptophyte algae (Tengs et al. 2000). In fact the optimal *psbB* tree embeds the peridinin dinoflagellates within the terminal clade of haptophytes, whereas the rRNA tree with 19' dinoflagellates places the anomalously pigmented dinoflagellates at the base of the haptophyte clade. However, the difference between these two trees (i.e. items 1 and 2 or 3; Table V.3) is, in the case of the *psbB* data, due to an inability to discriminate between relatively equally likely trees.

Moreover, the nuclear phylogeny of dinoflagellates suggests that this 19' clade evolved separately, and that these plastids do not represent the primitive condition for dinoflagellates (Litaker et al. 1999). Replacement of a peridinin-pigmented plastid with other plastids is a common theme in dinoflagellate evolution (Chesnick et al. 1996; Saldarriaga et al. 2001). Although the nuclear SSU rRNA phylogeny of dinoflagellates

may be incorrect, particularly given the generally low nonparametric bootstrap support seen in these trees (Saunders et al. 1997; Gunderson et al. 1999; Litaker et al. 1999; Saldarriaga et al. 2001), it is likely that the 19' clade of dinoflagellates, like other anomalously pigmented dinoflagellates, lost a peridinin type plastid and replaced it with one acquired from haptophytes.

However, the idea that the 19' containing dinoflagellates are the ancestral condition for dinoflagellates was supported by a recent study. This result may be artifactual considering the extreme branch lengths involved and it is unclear if this tree is substantially more likely than a tree separating the 19' and peridinin-containing dinoflagellates (Yoon et al. 2002). In fact the relationship of the 19' and peridinin-containing dinoflagellates is sensitive to the model of sequence evolution. When Bayesian analysis is performed with a model that allows the invariant sites parameter to vary the 19' and peridinin-containing dinoflagellates branch from different places within the haptophytes (Shalchian-Tabrizi 2003).

The *psbB* tree implies that the peridinin plastid of dinoflagellates was also derived from that of haptophytes. This would in turn imply that photosynthetic dinoflagellates evolved after photosynthetic haptophytes. The fossil record is difficult to interpret since the date for the earliest dinoflagellates varies enormously, and it is not clear if these early dinoflagellates were indeed photosynthetic. The fossil record of both dinoflagellates and haptophytes becomes more diverse in the Triassic, but there is enough ambiguity that it is possible that photosynthesis in both groups evolved more or less simultaneously (Tappan 1980). It is important to note that given the approximately unbiased test results our data cannot exclude the possibility that haptophytes have acquired their plastids from

dinoflagellates, or that they have both acquired their plastids from a similar unsampled source. Gibbs (1978) suggested that dinoflagellates acquired plastids from haptophytes because their plastids lack girdle lamellae and have thylakoids in stacks of three. Neither of these characters are unequivocal synapomorphies for these plastids, but they are ultrastructural features traditionally viewed as informative characters, and do seem to unite these groups.

#### *4.3. Models of chloroplast origin*

Almost all chromophytes use chlorophyll *c* and xanthophylls to harvest light (Bjørnland and Liaaen-Jensen 1989; Jeffrey 1989). This distinguishes them from the known red algae, which use phycobilins as their primary antenna pigments (Gantt 1981). This pigmentation character is a useful synapomorphy for chromophyte plastids, with the only possible intermediate form being the presence of both chlorophyll *c* and phycobilins in cryptomonads. If chromophyte plastids arose from several independent endosymbiotic events then they may have independently acquired similar pigmentation (the parallel model). Alternatively it is possible that the characteristic chromophyte pigmentation appeared in a lineage that acquired its plastids from red algae, and then later a second lineage acquired plastids from this first chromophyte lineage (the serial model). In an extreme form a serial model could propose, for example, that plastids were passed from red algae to cryptophytes, and thence to heterokonts, to haptophytes, and finally to dinoflagellates. At least some measure of serial plastid transfer seems to have occurred, as evidenced by the 19' clade of dinoflagellates (Tengs et al. 2000). The peridinin plastid could have arisen either from the same lineage as the haptophyte plastid in parallel, or could have been acquired from haptophytes in series and the phylogenetic signal would

be similar. Also, our data are not inconsistent with the idea that haptophytes could have acquired their plastids from dinoflagellates. Given the long branches of the dinoflagellates and our relatively poor understanding dinoflagellate and haptophyte diversity (Moon-van der Staay et al. 2001) we urge caution in literally interpreting these results. An integrated body of data will be required to accurately reconstruct the history of these organelles.

## **V.5.2. Materials and Methods**

### *Culture conditions*

The algae were cultured in *f/2* media without Si at 20°C with a 14hr/10hr L:D cycle at 24  $\mu\text{mol photons/m}^2\cdot\text{s}$  (Andersen et al. 1997). *Akashiwo sanguinea* (*Gymnodinium sanguineum*) cultures were obtained from Dr. Wayne Coats of the Smithsonian Environmental Research Center (SERC), and other dinoflagellate cultures were purchased from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP, see Table 1 for culture numbers). Haptophyte and heterokont cultures were obtained from the Marine Botany culture collection at the University of Oslo in Norway (MBUO).

### *2.2. Isolation of the A. sanguinea psbB*

Plastid DNA isolation from *A. sanguinea* was performed by isopycnic ultracentrifugation on CsCl gradients with Hoescht 33258, as described by Kite et al. (Kite et al. 1988). Briefly, total DNA was isolated with a CTAB buffer (2% CTAB, 0.7 M NaCl, EDTA), equilibrated first with equal volumes of phenol/chloroform, then with chloroform, and subsequently ethanol precipitated. The DNA was resuspended in TE (10 mM Tris pH 8.0 and 1mM EDTA) and combined with a CsCl solution to a final density

of 1.65 g/mL to which Hoescht 33258 had been added (200 µg/mL). This solution was centrifuged for 24-48 hrs in a Ti55 rotor at 30,000 rpm. The least dense band was pooled from several samples and re-isolated on a new CsCl gradient. This DNA was precipitated, restricted with EcoRI and ligated to EcoRI digested pBluescript SKm (Stratagene). From this library several clones were selected for sequencing with M13 primers using Big Dye (Applied Biosystems) terminators on the ABI 377. A 2.5kb clone was identified that contained the entire coding region of *psbB* from *A. sanguinea*.

### 2.3. PCR Amplification

PCR primers were designed from the novel sequence of *A. sanguinea* and the published sequence from *Heterocapsa triquetra*. The primer sequences are shown in Table V.2. Amplification was performed with a Biometra T-gradient thermocycler using an initial 30 second, 94°C denaturation followed by 35 cycles of 45-55°C annealing for 10 seconds, 72°C synthesis for 60-90 seconds, and 94°C denaturation for 10 seconds. Following amplification, samples were stored at 4°C until purified by polyethylene glycol precipitation (Morgan and Soltis 1995) and sequenced as described above.

### 2.4. Phylogenetic analysis

The sequences were edited with Sequencher 3.0 and manually aligned using MacClade 4.0 (Maddison and Maddison 2000). Phylogenetic analysis was performed with PAUP\* 4b4-10 (Swofford 2002). Parsimony searches were performed with 10 random addition replicates. For likelihood searches parameters were estimated from a Fitch-Margoliash distance tree generated with 10 addition sequence replicates. The initial distance measure was maximum likelihood using simultaneous estimates of the General Time Reversible (GTR) parameters and the empirical base frequencies. Because

**Table V.2 Sources of algal tissue and DNA sequences**

Classification	Source/Strain#	GenBank
<b>Dinoflagellates</b>		
<i>Akashiwo sanguinea</i>	SERC	new
<i>Ceratocorys horrida</i>	CCMP157	new
<i>Gymnodinium simplex</i>	CCMP419	new
<i>Katodinium rotundatum</i>	CCMP1542	new
<i>Gyrodinium impudicum</i>	CCMP1678	new
<i>Amphidinium operculatum</i>		<u>AJ250263</u>
<i>Heterocapsa triquetra</i>		<u>AF130034</u>
<b>Haptophytes</b>		
<i>Isochrysis galbana</i>	MBUO	<u>AJ575579</u>
<i>Prymnesium parvum</i>	MBUO	<u>AJ575581</u>
<i>Emiliana huxleyi</i>	MBUO	<u>AJ575578</u>
<i>Pavlova gyrans</i>	MBUO	<u>AJ575580</u>
<b>Heterokonts</b>		
<i>Odontella sinensis</i>		<u>1185127</u>
<i>Pelagomonas calceolata</i>	CCMP 1214	new
<i>Chattonella sp.</i>	MBUO	new
<b>Cryptophytes</b>		
<i>Guillardia theta</i>		<u>AF041468</u>
<b>Rhodophytes</b>		
<i>Porphyra purpurpea</i>		<u>U38804</u>
<i>Cyanidium caldarium</i>		<u>AF022186</u>
<b>Chlorophytes</b>		
<i>Nicotiana tabacum</i>		<u>Z0004</u>
<i>Marchantia polymorpha</i>		<u>11640</u>
<i>Chlorella vulgaris</i>		<u>2224352</u>
<b>Euglenozoa</b>		
<i>Euglena gracilis</i>		<u>415327</u>
<b>Glaucophyte</b>		
<i>Cyanophora paradoxa</i>		<u>U30821</u>
<b>Cyanobacteria</b>		
<i>Synechocystis sp.</i>	PCC6803	<u>1001200</u>
<i>Anabaena (Nostoc)</i>	PCC7120	<u>X58847</u>

**Table V.3 Primers used for amplification and sequencing of *psbB***

Primer name	primer sequence
<b>Haptophytes / Heterokonts</b>	
F1	ATGGCGTTACCATGGTATCGTGT
F129	GTGGCGTCAGGGTATGTTTGT
F500	TTGGTCCTGGTATTTGGGT
F1200	GCTGTAACTTCTATGGTGG
R500	ACCCAAATACCAGGACCAA
R1200	CCACCATAGAAGTTACAGC
R1508	TTACACAGCACCTTGTTTCTTAG
R1500s	TCTAAACTTCAGCACCAATACC
<b>Dinoflagellates</b>	
PsbB 495R	GAGAGCACGAGCTCCATGCCACAA
PsbB 500R	AAGAGAGCACGAGCTCCATGCCA
PsbB -972F	CCTTGGTTTATAGGGTACATATCGT
PsbB -90R	GTCCATCTCGATATCACTGGGA
PsbB -90F	CAGTGATATCGAGATGGACCAAA
Gsimplex psbBF	GGCATATTAGCTCAAGACCTGGA
Gsimplex psbBR	ACAACTGTTGTTGCACTGCCA

site to site rate parameters cannot be simultaneously estimated while constructing distance matrices, the invariant site parameter was arbitrarily set to zero and the gamma distribution parameter was set to 0.5 with four categories. The resulting tree was used to re-estimate all of the likelihood parameters including invariant sites and gamma. These parameters were used to construct another likelihood distance matrix, and another Fitch-Margoliash tree. The parameters were then re-estimated from this final distance tree, and used for all subsequent searches. The likelihood ratio test was used to evaluate alternate models of sequence evolution. Nonparametric bootstrapping was performed on 100 resampled datasets using the same model parameters used to derive the optimum tree.

For parametric bootstrapping 100 datasets were simulated (Rambaut and Grassly 1997) by SeqGen V1.1 based on the optimum tree topology, branch lengths, and likelihood model (Fig. V.2). These simulated datasets were then analyzed using PAUP with the same ML model parameters. Because SeqGen does not use invariant site rate correction a gamma correction with five categories was used both to simulate the sequences and to construct the optimum tree using the sequences. Two different trees were used to simulate sequences, the parsimony tree (data not shown) and the likelihood tree.

For Bayesian analysis MrBayes (Huelsenbeck and Ronquist 2001) was used with the following permutations: a single set of model parameters for all three codon positions, one set of parameters for each position, a single model using only the first two positions, and two sets of parameters one for each of the remaining positions. Additionally the covarion option was used with several of these combinations. In all cases the four

Markov chains were run for  $2 \times 10^6$  generations with one chain being heated, the burnin period was 10,000 generations and trees were sampled every one hundredth generation. Another set of analyses were performed using PAUP, but excluding the third codon position. The all position tree was used as a starting tree and parameters were estimated simultaneously with the heuristic search. PAML searches were attempted on a translated version of the nucleotide dataset, **using random addition and resolved** starting trees. For the approximately unbiased test using this dataset fourteen alternate hypotheses were constructed. For ten of these an optimal tree was constructed using PAUP and one or more constraints the other four hypotheses represent potentially optimal trees using different methods and models (detailed in Table V.3). In each of these cases the GTR + I +  $\Gamma$  parameters derived from the heuristic search were used. The PAML (Yang 1997) package was also used to get site likelihoods of these trees, and the optimal trees from the PAUP analysis were used with the Wagner matrix and the shape parameter set to 1.03. In addition sitelikelihoods for these same trees were used when the third codon position was excluded. The site likelihoods for all fourteen trees were then compared using CONSEL (Shimodaira 2000; Shimodaira and Hasegawa 2001).

## Chapter VI Conclusions

### VI.1. The mystery of dinoflagellate chloroplast DNA

For many years dinoflagellate chloroplast DNA was a mysterious substance. Early work established that some DNA was there (Franker 1970; Kowallik and Haberkorn 1971), but after no restriction maps or Southern blots appeared for dinoflagellates it seemed that work on these organisms was falling behind. A clue to the odd nature of this plastid could have been the ferredoxin (*petF*) amino acid sequence (Uchida et al. 1988). However, as other chloroplast genomes were surrendering their mysteries to chloroplast DNA purification or polymerase chain reaction techniques dinoflagellates remained unknown. Some reasons for this are clear. Dinoflagellates are difficult to culture, and bacterial contamination is expected.

Added to the difficulty in culturing dinoflagellates is the remarkably skewed DNA content of dinoflagellate cells. Because dinoflagellate nuclei have large DNA contents and the minicircles are relatively small there is an extreme bias against the chloroplast DNA. Dinoflagellate minicircle DNA was eventually purified independently in several labs, but few genes were found (Zhang et al. 1999; Barbrook and Howe 2000). Chloroplast genes were also simultaneously independently amplified using PCR and reverse transcription techniques (Takishita and Uchida 1999). However, the extreme divergence of the dinoflagellate chloroplast genes prevented rapid sequence acquisition using PCR.

By using an EST approach many of these problems were circumvented, but again, predictably, cDNA library construction was relatively difficult. Because only expressed portions of the genome were sampled there was a relatively good chance that some highly

expressed genes would be chloroplast targeted. Fortunately many highly expressed genes were found that are homologous to typically chloroplast-encoded genes. This fits very well with the reduced coding potential of the minicircle-encoded genes.

The combination of the minicircle and EST data imply that the eleven protein-coding genes that have been found in dinoflagellate minicircles may be the only plastid-encoded genes in dinoflagellates. If this inference is correct then why and how could gene transfer on this scale have happened? As was shown in chapter IV, part of the reason for gene transfer might be to rescue genes from the apparently high rate of mutation in the chloroplast. One concept could be that dinoflagellates have a very efficient mechanism for moving genes into the nucleus. The paucity of mitochondrial data suggests that dinoflagellate mitochondria also have relatively few protein-coding genes. This idea is corroborated by EST data. Three mitochondrial genes were relatively common in both libraries. The A+T bias in the *Gonyaulax* putative mitochondrial sequences suggests a non-nuclear origin. Clearly more work is needed to confirm these mitochondrial data, but it seems possible that the coding potential of the mitochondrion is reduced, like the coding potential of the chloroplast.

Is the organelle to nucleus gene transfer seen in dinoflagellates a symptom of an underlying gene transfer mechanism? Could this gene mobilization mechanism allow dinoflagellates to enter into endosymbiotic relationships more easily? Several hypotheses can be tested that could expose this mechanism. For example the degree of chloroplast to nucleus gene transfer and nucleus-to-nucleus gene transfer could be assessed in the haptophyte-containing dinoflagellates, or the diatom-containing *Peridinium foliaceum*. Even relatively small-scale genomic data from *Amphidinium* could indicate the number

of copies of typically plastid-encoded genes in the genome, and perhaps reveal common insertion sites in the chromosomal DNA. Instead of a random sequencing approach, a directed sequencing effort using known plastid targeted genes could be used.

Dinoflagellate chloroplasts may use genes from many different lineages (mosaic evolution) as is implied by some of the gene trees presented in the previous chapter. Sub-genomic sequencing might reveal additional suspicious genes.

## **VI.2.Phylogeny: some testable hypotheses**

Given present taxon sampling and phylogenetic methods, is the mosaic hypothesis testable? The data presented in Chapter III can not reject this hypothesis, but also cannot support this hypothesis. For example the *atpG* gene tree implies a green algal ancestry, but this result does not have enough bootstrap support. Even if the bootstrap support were high phylogenetic artifact cannot be ruled out. For example nucleotide bias is very different between the chloroplast (typically A+T rich) and the nucleus. So, for genes that are nuclear-encoded in two unrelated lineages there could be a significant chance of convergence. But shared nucleotide composition could also indicate descent from a common ancestor. Worse, how can the significance of nucleotide, or amino acid bias be judged?

One way to better understand the known data would be to acquire more data. For example genes that are known to be nuclear-encoded in all plastid lineages could be sampled. This might help overcome the difficulty in cross compartment phylogeny. One example of this category of gene would be the LHC genes. The preliminary data from *Amphidinium* suggests broad diversity in this family, so it may be a difficult phylogenetic marker. Other nuclear-encoded plastid-targeted genes might include *psbO*, that has

already been sampled from dinoflagellates and haptophytes (Ishida and Green 2002). Specifically new genes with finer taxon sampling could be used to test the hypotheses of a haptophyte or heterokont origin of the dinoflagellate plastid. Certainly much finer taxon sampling would be needed to nest either the haptophytes or dinoflagellates within the heterokonts.

The analyses presented in Chapters III and V are difficult to interpret, in the case of minicircle genes because of extreme branch lengths. The nuclear-encoded genes are difficult largely because they are nuclear-encoded and subject to a very different selection regime. One way to interpret these results would be to introduce a trichotomy between dinoflagellates, heterokonts and haptophytes and by necessity exclude cryptophytes.

Figure VI.1 shows an evolutionary scheme where some chloroplast characters are mapped onto a phylogenetic tree. Although more data is needed to make a really well supported case for cryptophytes as the chromophyte outgroup there are ultrastructural and biochemical features that make cryptophytes a likely outgroup to the other chlorophyll *c* containing plastids.

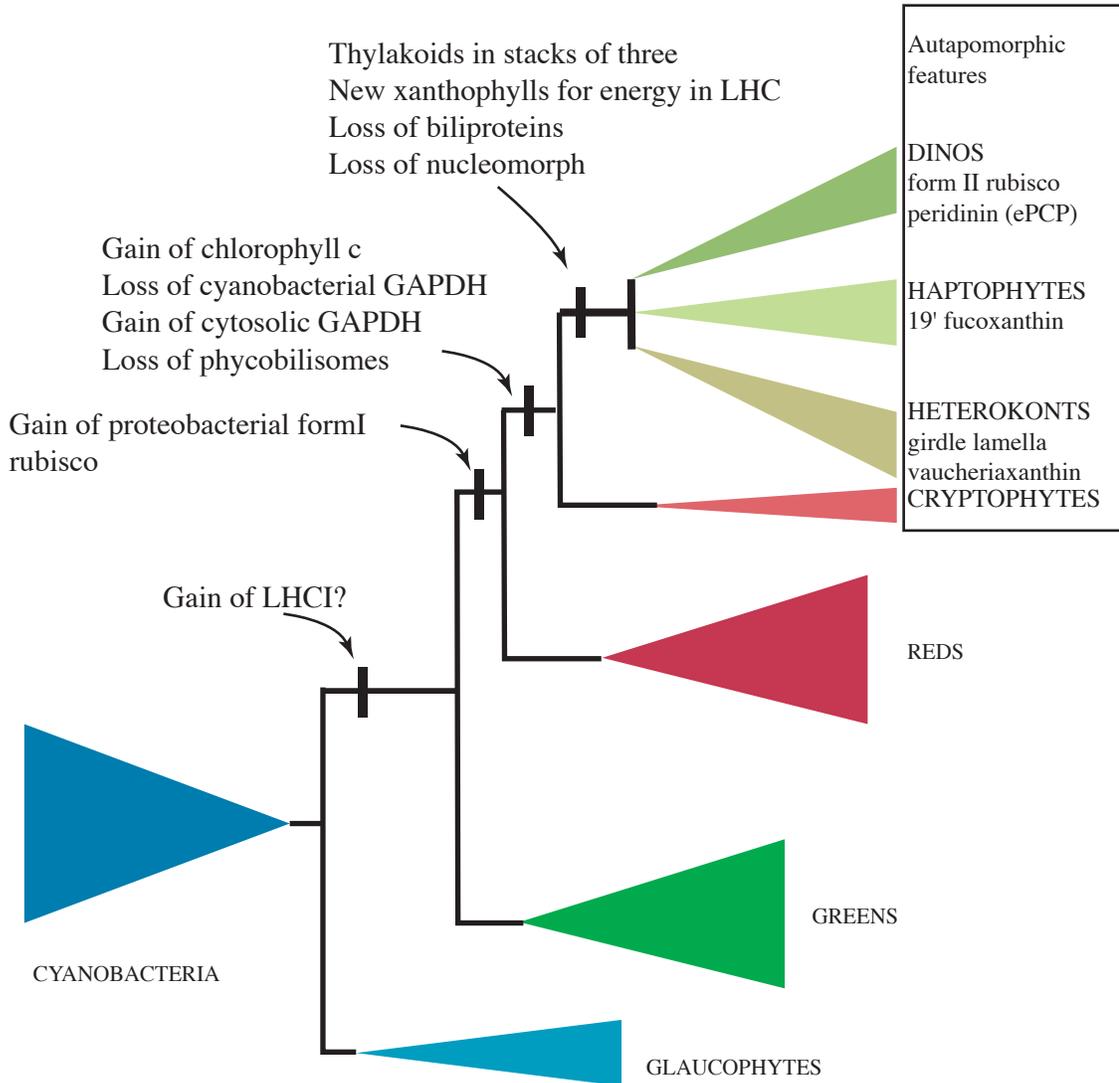
Cryptophytes are the only chromophytes to retain the red algal nucleus, and also retain phycobilipigments. These pigments are not arranged in typical red algal or cyanobacterial phycobilisomes, and in fact are found in the intrathylakoidal spaces (Gantt et al. 1971). These two characters, a nucleomorph and vestigial red algal pigment are readily interpreted as ancestral. This interpretation does not guarantee that other characters of the cryptophytes are necessarily ancestral.

In a biosynthetic scheme Bjørnland and Liaaen-Jensen (1989) suggest that the cryptophyte light-harvesting xanthophyll alloxanthin (Fig. I.3) could be a precursor to

fucoxanthin (Fig. I.4) and peridinin (Fig. I.1). One major problem to be worked out is whether the light-harvesting complexes of chromophytes are more pigment specific than the light-harvesting complexes of red algae. Red algal complexes can be reconstituted with pigments from several different lineages (Grabowski et al. 2001). If this result is extrapolated to chromophyte light-harvesting complexes then it seems likely that chlorophyll and carotenoid binding is specified by other factors such as biosynthesis.

Understanding the evolution of these xanthophylls and of chlorophyll *c* is based on increasing resolution of chromophyte phylogeny. Chlorophyll *c* biosynthesis is a major innovation that maps cleanly onto the phylogeny presented here. The predicted pathway of chlorophyll *c* biosynthesis suggests that a single side reaction in the biosynthesis of chlorophyll *a* could produce chlorophyll *c* (Jeffrey 1989). Perhaps an unidentified EST for chlorophyll *c* biosynthesis has already been sequenced, and biochemical techniques could be used to uncover this gene. The same genes that specify chlorophyll *c* biosynthesis and carotenoid biosynthesis might in turn provide further resolution of plastid phylogeny.

Figure VI.1 A phylogenetic reconstruction of chloroplast evolution.



## REFERENCES

- Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie and J. C. Venter (1991). Complementary-DNA sequencing - expressed sequence tags and human genome project. *Science* **252**: 1651-1656.
- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389-3402.
- Andersen, R. A., S. L. Morton and J. P. Sexton (1997). CCMP - Provasoli-Guillard National Center for Culture of Marine Phytoplankton. *J. Phycol.* **33**: Supplement.
- Anderson, D. M. (1994). Red Tides. *Scientific American* **271**: 62-8.
- Anderson, D. M., A.D. Cemballa and G. M. Hallegraeff. (1998). *Physiological Ecology of Harmful Algal Blooms*. New York, Springer.
- Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **393**: 796-815.

- Bachvaroff, T. R., G. T. Concepcion, C. R. Rogers and C. F. Delwiche (2004).  
Dinoflagellate EST data indicate massive transfer of chloroplast genes to the  
nucleus. *Protist* **155**: 65-78.
- Baldauf, S. (2003). The deep roots of eukaryotes. *Science* **5626**: 1703-1706.
- Baldauf, S. L., A. J. Roger, I. Wenk-Siefert and W. F. Doolittle (2000). A kingdom-level  
phylogeny of eukaryotes based on combined protein data. *Science* **290**: 972-977.
- Barbrook, A. C. and C. J. Howe (2000). Minicircular plastid DNA in the dinoflagellate  
*Amphidinium operculatum*. *Mol Gen Genet* **263**: 152-158.
- Barbrook, A. C., H. Symington, R. E. Nisbet, A. Larkum and C. J. Howe (2001).  
Organisation and expression of the plastid genome of the dinoflagellate  
*Amphidinium operculatum*. *Mol Gen Genet* **266**: 632-8.
- Bjørnland, T. and S. Liaaen-Jensen (1989). Distribution patterns of carotenoids in  
relation to chromophyte phylogeny and systematics. *The Chromophyte algae:  
Problems and Perspectives*. J. C. Green, B. S. C. Leadbeater and W. L. Diver  
(eds). Oxford, Clarendon Press. **38**: 37-62.
- Bjørnland, T. and K. Tangen (1979). Pigmentation and morphology of a marine  
*Gyrodinium* (Dinophyceae) with a major carotenoid different from peridinin and  
fucoxanthin. *J Phycol* **15**: 457-463.

- Bockstahler, K. R. and D. W. Coats (1993). Spatial and temporal aspects of mixotrophy in Chesapeake Bay dinoflagellates. *J. Euk. Microb.* **40**: 49-60.
- Bricker, T. M. (1990). The structure and function of CPa-1 and CPa-2 in Photosystem II. *Photosynthesis Res.* **21**: 1-13.
- Buchanan, B. B., W. Gruissem and R. L. Jones (2000). *Biochemistry and Molecular Biology of Plants*. Rockville, MD, American Society of Plant Physiologists.
- Burger, G., D. Saint-Louis, M. W. Gray and B. F. Lang (1999). Complete sequence of the mitochondrial DNA of the red algae *Porphyra purpurea* shared cyanobacterial introns and shared ancestry of red and green algae. *Plant Cell* **11**: 1675-1694.
- Cachon, J., M. Cachon and P. Salvano (1979). The nuclear division of *Oxyrrhis marina*: an example of the role played by the nuclear envelope in chromosome segregation. *Arch. Protistenkd* **122**: 43-54.
- Cachon, J., M. Cachon (1987). Parasitic Dinoflagellates. *The Biology of Dinoflagellates*. F. J. R. Taylor (ed). Oxford, Blackwell Scientific Publications. **21**: 571-610.
- Cavalier-Smith, T. (1993). Kingdom Protozoa and its 18 Phyla. *Microbiological Reviews* **57**: 953-994.

- Cavalier-Smith, T. (1999). Principles of Protein and Lipid Targeting in Secondary Symbiogenesis: Euglenoid, Dinoflagellate, and Sporozoan Plastid Origins and the Eukaryote Family Tree. *J. Eukaryot. Microbiol.* **46**: 347-366.
- Cavalier-Smith, T. (2000). Membrane heredity and early chloroplast evolution. *Trends in Plant Science* **5**: 174-182.
- Cavalier-Smith, T., M. T. E. P. Allsopp and E. E. Chao (1994). Chimeric conundra: Are nucleomorphs and chromists monophyletic or polyphyletic? *Proc. Natl. Acad. Sci USA* **91**: 11368-11372.
- Chaput, H., Y. Wang and D. Morse (2002). Polyadenylated transcripts containing random gene fragments are expressed in dinoflagellate mitochondria. *Protist.* **153**: 111-22.
- Chatton, E. (1920). Les peridiniens parasites. *Archives de zoologie experimentale et generale* **59**: 1-475.
- Chesnick, J. M. and R. Cattolico (1993). Isolation of DNA from eukaryotic algae. *Method Enzymol* **224**: 168-176.
- Chesnick, J. M., C. W. Morden and A. M. Schmeig (1996). Identity of the endosymbiont of *Peridinium foliaceum* (Pyrrophyta): analysis of the *rbcLS* operon. *J. Phycol.* **32**: 850-857.

- Chomczynski, P. and N. Sacchi (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical biochemistry*. **162**: 156-9.
- Christensen, T. (1962). *Alger. Systematisk Botanik*. T. W. Bocher and M. Lange (eds). Copenhagen, Munksgaard. **2**.
- Christensen, T. (1989). The Chromophyta, past and present. *The Chromophyte Algae: Problems and Perspectives*. J. C. Green, B. S. C. Leadbeater, W. L. Diver (eds). Oxford, Clarendon Press. **38**: 1-12.
- Coats, D. W. (1999). Parasitic life styles of marine dinoflagellates. *J of Euk. Microbiol* **46**: 402-409.
- Daugbjerg, N. and R. A. Andersen (1997). Phylogenetic analyses of the *rbcL* sequences from haptophytes and heterokont algae suggest their chloroplasts are unrelated. *Mol. Biol. Evol* **14**: 1242-51.
- Daugbjerg, N., G. Hansen, J. Larsen and O. Moestrup (2000). Phylogeny of some of the major genera of dinoflagellates based on ultrastructure and partial LSU rDNA sequence data, including the erection of three new genera of unarmoured dinoflagellates. *Phycologia* **39**: 302-317.

Delwiche, C. F. (1999). Tracing the thread of plastid diversity through the tapestry of life.  
*American Naturalist* **154(S4)**: 164-177.

Delwiche, C. F., M. Kuhsel and J. D. Palmer (1995). Phylogenetic analysis of *tufA*  
sequences indicates a cyanobacterial origin of all plastids. *Molecular*  
*phylogenetics and evolution*. **4**: 110-28.

Delwiche, C. F. and J. D. Palmer (1996). Rampant horizontal transfer and duplication of  
rubisco genes in eubacteria and plastids. *Mol Biol and Evol* **13**: 873-882.

Delwiche, C. F. and J. D. Palmer (1997). The origin of plastids and their spread via  
secondary endosymbiosis. *Plant Systematics and Evolution [Supplement]* **11**: 53-  
86.

Dodge, J. D. (1965). Chromosome structure in the dinoflagellates and the problem of the  
mesokaryotic cell. *Int. Congr. Ser. -- Excerpta Med.* **91**: 339.

Dodge, J. D. (1966). *The Dinophyceae. The Chromosomes of the Algae*. M. B. E.  
Godward (ed). New York, St. Martin's press.

Dodge, J. D. (1971). A dinoflagellate with both a mesokaryotic and eukaryotic nucleus.  
I. Fine structure of the nuclei. *Protoplasma* **73**: 145-157.

- Dodge, J. D. (1975). A survey of chloroplast ultrastructure in the Dinophyceae. *Phycologia* **14**: 253-263.
- Dodge, J. D. and C. Greuet (1987). Dinoflagellate ultrastructure and complex organelles. *The Biology of Dinoflagellates*. F. J. R. Taylor (ed). Oxford, UK, Blackwell Scientific Publications. pp 92-142.
- Douady, C. J., F. Catzefflis, J. Raman, M. S. Springer and M. J. Stanhope (2003). The Sahara as a vicariant agent, and the role of Miocene climatic events, in the diversification of the mammalian order Macroscelidea (elephant shrews). *Proc. Natl. Acad. Sci USA* **100**: 8325-8330.
- Douglas, S., Zauner S., Fraunholz M., Beaton M., Penny S., Deng L. T., Wu X., Reith M., Cavalier-Smith T., Maier U.G. (2001). The highly reduced genome of an enslaved algal nucleus. *Nature* **410**: 1091-1096.
- Douglas, S. E. and S. L. Penny (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.
- Douglas, S. E. and S. L. Penny (1999). The plastid genome of the cryptophyte alga, *Gullardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J. Mol. Evol.* **48**: 236-244.

- Durnford, D. G., J. Dean, S. Tan, G. I. McFadden, E. Gantt and B. R. Green (1999). Phylogenetic assessment of the eukaryotic light-harvesting antenna proteins, with implications for plastid evolution. *J. Mol. Evol.* **48**: 59-68.
- Eaton, G. (1980). Nomenclature and homology in perdinialean dinoflagellate plate pattern. *Paleontology* **23**: 667-688.
- Egea, N. and N. Lang-Unnasch (1995). Phylogeny of the large extrachromosomal DNA of organisms in the phylum Apicomplexa. *J. Eukaryot. Microbiol.* **42**: 679-684.
- Emanuelsson, O., H. Nielsen and G. von Heijne (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science* **8**: 978-984.
- Ewing, B., L. Hillier, M. C. Wendl and P. Green (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175-185.
- Fagan, T., J. W. Hastings and D. Morse (1998). The phylogeny of glyceraldehyde-3-phosphate dehydrogenase indicates lateral gene transfer from cryptomonads to dinoflagellates. *J Mol Evol* **47**: 633-639.
- Fast, N. M., J. C. Kissinger, D. S. Roos and P. J. Keeling (2001). Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol Biol Evol* **18**: 418-426.

- Felsenstein, J. (1978). Cases in which parsimony and compatibility methods will be positively misleading. *Cladistics* **10**: 315-319.
- Fichera, M. E. and D. S. Roos (1997). A plastid organelle as a drug target in apicomplexan parasites. *Nature* **390**: 407-409.
- Franker, C. (1970). Some properties of DNA from zooxanthellae harbored by an anemone *Anthopleura elegantissima*. *J Phycol* **6**: 299-305.
- Gaines, G. and M. Elbrächter (1987). Heterotrophic nutrition. The biology of dinoflagellates. F. J. R. Taylor (ed). Oxford, Blackwell pp. 224-268.
- Gajadhar, A. A., W. C. Marquardt, R. Hall, J. Gunderson, E. V. Ariztia-Carmona and M. L. Sogin (1991). Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Cryptosporidium parvum* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. *Molecular and Biochemical Parasitology* **45**: 147-154.
- Gantt, E. (1981). Phycobilisomes. *Annual Review of Plant Physiology* **32**: 327-347.
- Gantt, E., M. R. Edwards and L. Provosoli (1971). Chloroplast structure of the Cryptophyceae. *J Cell Biol.* **48**: 280-290.

Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser and B. Barrell (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**: 498-511.

Gardner, M. J., H. Tettelin, D. J. Carucci, L. M. Cummings, L. Aravind, E. V. Koonin, S. Shallom, T. Mason, K. Yu, C. Fujii, J. Pederson, K. Shen, J. Jing, C. Aston, Z. Lai, D. C. Schwartz, M. Pertea, S. Salzberg, L. Zhou, G. G. Sutton, R. Clayton, O. White, H. O. Smith, C. M. Fraser, S. L. Hoffman and et al. (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science*. **282**: 1126-32.

Gibbs, S. P. (1962). Nuclear envelope-chloroplast relationships in algae. *J Cell Biol* **14**: 433-444.

Gibbs, S. P. (1978). The chloroplast of *Euglena* may have evolved from symbiotic green algae. *Can. J. Bot* **56**: 2883-2889.

- Gibbs, S. P. (1981). The chloroplast endoplasmic reticulum, structure, function, and evolutionary significance. *International Review of Cytology* **72**: 49-99.
- Gibbs, S. P. (1993). The evolution of algal chloroplasts. *Origins of Plastids*. R. A. Lewin (ed), Chapman and Hall, New York. pp. 107-121.
- Gillot, M. A., and S. P. Gibbs (1980). The Cryptomonad nucleomorph: Its ultrastructure and evolutionary significance. *J. Phycol.* **16**: 558-568.
- Glockner, G., A. Rosenthal and K. Valentin (2000). The structure and gene repertoire of an ancient red algal plastid genome. *J Mol Evol* **51**: 382-390.
- Grabowski, B., F. X. Cunningham and E. Gantt (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, L. E. and L. W. Wilcox (2000). *Algae*. Upper Saddle River, NJ, Prentice-Hall.
- Gray, M. W. and D. F. Spencer (1996). Organellar evolution. *Evolution of Microbial Life*. D. McL. Roberts, P. Sharp, G. Alderson and M. Collins (eds). Cambridge, Cambridge University Press: 109-126.

- Grzebyk, D., O. Schofield, C. Vetriani and P. G. Falkowski (2003). The mesozoic radiation of eukaryotic algae: the portable plastid hypothesis. *J Phycol* **39**: 259-267.
- Gunderson, J. H., S. H. Goss and D. W. Coats (1999). The phylogenetic position of *Amoebophrya* sp. from *Gymnodinium sanguineum*. *J. Euk. Microbiol.*, **46**: 194-197.
- Gustafson, D. E. J., D.K. Stoecker, M.D. Johnson, W.F. Van Heukelem, and K. Sneider (2000). Cryptophyte algae are robbed of their organelles by the marine ciliate *Mesodinium rubrum*. *Nature* **405**: 1049-1052.
- Hackett, J. D., H. S. Yoon and D. Bhattacharya (2003). Evidence for the cryptophyte origin of the plastid of *Dinophysis* (Dinophysiales, Dinophyceae). *J Phycol.* **39**: 440-448.
- Hackett, J. D., H. S. Yoon, M. B. Soares, M. F. Bonaldo, T. L. Casavant, T. E. Scheetz, T. Nosenko and D. Bhattacharya (2004). Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. *Current Biology* **14**: 213-218.
- Harper, J. T. and P. J. Keeling (2003). Nucleus-encoded, plastid-targeted glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. *Mol Biol Evol* **20**: 1730-1735.

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

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

Hiller, R. G. (2001). 'Empty' minicircles and *petB/atpA* and *psbD/psbE* (*cytb<sub>559</sub>*  $\alpha$ ) genes in tandem in *Amphidinium carterae* plastid DNA. FEBS letters **505**: 449-452.

Hiller, R. G., P. M. Wrench and F. P. Sharples (1995). The light harvesting chlorophyll *a*-*c*-binding protein of dinoflagellates: a putative polyprotein. FEBS letters **363**: 175-178.

Hillis, D. M. and J. J. Bull (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42**: 182-192.

Hofmann, E., P.M. Wrench, F.P. Sharples, R.G. Hiller, W. Welte and K. Diederich. (1996). Structural basis of light harvesting by carotenoids: peridinin-chlorophyll-protein from *Amphidinium carterae*. Science **272**: 1788-1791.

Huelsenbeck, J. P. and B. Rannala (1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science. **276**: 227-232.

- Huelsenbeck, J. P., D. M. Hillis and R. Jones (1996). Parametric bootstrapping in molecular phylogenetics: applications and performance. *Molecular Zoology: Advances, Strategies, and Protocols*. J. D. Ferraris and S. R. Palumbi (eds), Wiley-Liss Inc.
- Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**: 754-755.
- Ishida, K., T. Cavalier-Smith and B. R. Green (2000). Endomembrane structure and the chloroplast protein targeting pathway in *Heterosigma akashiwo* (Raphidophyceae, Chromista). *J of Phycol* **36**: 1135-1144.
- Ishida, K. and B. R. Green (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.
- Jeffrey, S. W. (1989). Chlorophyll *c* pigments and their distribution in the chromophyte algae. *The chromophyte algae problems and perspectives*. J. C. Green, B. S. C. Leadbeater and W. L. Diver (eds). Oxford, Clarendon Press.

Jeffrey, S. W. and S. W. Wright (1994). Photosynthetic pigments in the Haptophyta. The Haptophyte algae. J. C. Green, B. S. C. Leadbeater (eds). Oxford, Clarendon Press. **51**: 111-132.

Jenks, A. and S. P. Gibbs (2000). Immunolocalization and distribution of form II rubisco in the pyrenoid and chloroplast stroma of *Amphidinium carterae* and form I rubisco in the symbiont-derived plastids of *Peridinium foliaceum* (dinophyceae). *J Phycol* **34**: 1009-1016.

Kaneko, T., Y. Nakamura, C. P. Wolk, T. Kuritz, S. Sasamoto, A. Watanabe, M. Iriguchi, A. Ishikawa, K. Kawashima, T. Kimura, Y. Kishida, M. Kohara, M. Matsumoto, A. Matsuno, A. Muraki, N. Nakazaki, S. Shimpo, M. Sugimoto, M. Takazawa, M. Yamada, M. Yasuda and S. Tabata (2001). Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* **8**: 205-13.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Kakazaki, K. Naruo, S. Okumura and S. S. , C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, and S. Tabita (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.

- Karol, K. G., R. M. McCourt, M. T. Cimino and C. F. Delwiche (2001). The closest living relatives of land plants. *Science* **294**: 2351-3.
- Kite, G. C., L. J. Rothschild and J. D. Dodge (1988). Nuclear and plastid DNAs from the binucleate dinoflagellates *Glenodinium (Peridinium) foliaceum* and *Peridinium balticum*. *Biosystems* **21**: 151-163.
- Kofoed, C. (1909). On *Peridinium steini* Jörgensen, with a note on the nomenclature of the skeleton of the Peridinidae. *Arch. Protistenkd* **16**: 26-47.
- Kohler, S., C.F. Delwiche, P.W. Denny, L.G. Tilney, P. Webster, R.J.M. Wilson, J.D. Palmer and D. S. Roos (1997). A plastid of probable green algal origin in apicomplexan parasites. *Science* **275**: 1485-1489.
- Kowallik, K. V. and G. Haberkorn (1971). The DNA-structures of the chloroplast of *Prorocentrum micans* (Dinophyceae). *Archiv für Mikrobiologie* **80**: 154-165.
- Kowallik, K. V., B. Stoebe, I. Schaffran, K.-P. P and U. Freier (1995). The chloroplast genome of a chlorophyll *a+c* containing alga, *Odontella sinensis*. *Plant Molecular Biology Reporter* **13**: 336-342.
- Kubai, D. F. and H. Ris (1969). Division in the dinoflagellate *Gyrodinium cohnii* (Schiller). A new type of nuclear reproduction. *J Cell Biol.* **40**: 508-28.

- Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray and G. Burger (2002). The closest unicellular relatives of animals. *Current Biology* **12**: 1773-1778.
- Lang, B. F., E. Seif, M. W. Gray, C. J. O'Kelly and G. Burger (1999). A comparative genomics approach to the evolution of eukaryotes and their plastids. *J Euk Microbiol* **46**: 320-326.
- Lemieux, C., C. Otis and M. Turmel (2000). Ancestral chloroplast genome in *Mesostigma virde* reveals an early branch of green plant evolution. *Nature* **403**: 649-652.
- Lenaers, G., C. Scholin, Y. Bhaud, D. Saint-Hilaire and M. Herzog (1991). A molecular phylogeny of dinoflagellate protists (Pyrrhophyta) inferred from the sequence of 24S rRNA divergent domains D1 and D8. *J. Mol. Evol.* **32**: 53-63.
- Liaud, M. F., U. Brandt, M. Scherzinger and R. Cerff (1997). Evolutionary origin of cryptomonad microalgae: two novel chloroplast/cytosol-specific GAPDH genes as potential markers of ancestral endosymbiont and host cell components. *J Mol Evol* **44 Suppl 1**.
- Litaker, R. W., P. A. Tester, A. Coloni, M. G. Levy and E. J. Noga (1999). The phylogenetic relationship of *Pfiesteria piscicida*, *Cryptoperidiniopsis* sp. *Amyloodinium ocellatum* and a *Pfiesteria*-like dinoflagellate to other dinoflagellates and apicomplexans. *J. Phycol.* **35**: 1379-1389.

- Loeblich, A. R. I. (1976). Dinoflagellate evolution: Speculation and evidence. *J. Protozool.* **23**: 13-28.
- Lopez-Garcia, P., F. Rodriguez-Valera, C. Pedros-Alio and D. Moriera (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**: 603-607.
- Lucas, I. A., M. Vesik (1990). The fine structure of two photosynthetic species of *Dinophysis* (Dinophysiales, Dinophyceae). *J. Phycol.* **26**: 345-357.
- Maddison, W. P. and P. R. Maddison (2000). MacClade version 4: Analysis of phylogeny and character evolution. Sunderland, MA, Sinauer Associates.
- Marchant, H. J. and H. A. Thomsen (1994). Haptophytes in polar waters. *The Haptophyte Algae*. J. C. Green and B. S. C. Leadbeater (eds). Oxford, Clarendon Press.
- Margulis, L. (1971). The origin of plant and animal cells. *Am Sci* **59**: 230-235.
- Marin, B., M. Klingberg and M. Melkonian (1998). Phylogenetic relationships among the Cryptophyta: Analyses of nuclear-encoded SSU rRNA sequences support the monophyly of extant plastid-containing lineages. *Protist* **149**: 265-276.
- Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelson, T. Lins, D. Leister, B. Stoebe, M. Hasegawa and D. Penny (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.
- Martin, W., B. Stoebe, V. Goremykin, S. Hansmann, M. Hasegawa and K. V. Kowallik (1998). Gene transfer to the nucleus and the evolution of chloroplasts. Nature **393**: 162-165.
- McFadden, G. I., P. R. Gilson and D. R. A. Hill (1994). *Goniomonas*: rRNA sequences indicate that this phagotrophic flagellate is a close relative of the host component of cryptomonads. Eur. J. Phycol. **29**: 29-32.
- Moon-van der Staay, S., R. De Wachter and D. Vaultot (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature **409**: 607-610.
- Morgan, D. R. and D. E. Soltis (1995). Phylogenetic relationships among members of Saxifragaceae *sensu lato* based on *rbcL* sequence data. Ann. Mo. Bot. Gar. **82**: 208-234.
- Morse, D., P. Salois, P. Markovic and J. W. Hastings (1995). A nuclear-encoded form II rubisco in dinoflagellates. Science **268**: 1622-1624.

- Nassoury, N., M. Cappadocia and D. Morse (2003). Plastid ultrastructure defines the protein import pathway in dinoflagellates. *Journal of Cell Science* **116**: 2867-2874.
- Netzel, H. and G. Dürr (1984). Dinoflagellate cell cortex. *Dinoflagellates*. D. L. Spector (ed). Orlando, Academic Press.
- Nielsen, H., J. Engelbrecht and S. Brunak (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**: 1-6.
- Nikaido, I., E. Asamizu, M. Nakajima, Y. Nakamura, N. Saga and S. Tabata (2000). Generation of 10,154 expressed sequence tags from a leafy gametophyte of a marine red alga, *Porphyra yezoensis*. *DNA Res.* **7**: 223-227.
- Noren, F., O. Moestrup and A.-S. Rehnstam-Holm (1999). *Parvilucifera infectans* gen. et sp. nov. (Perkinsozoa phylum nov.): A parasitic flagellate capable of killing toxic algae. *Eur. J. Protistol.* **35**: 233-54.
- Ohta, N., M. Matsuzaki, O. Misumi, S. Y. Miyagishima, H. Nozaki, K. Tanaka, T. Shin-I, Y. Kohara and T. Kuroiwa (2003). Complete sequence and analysis of the plastid genome of the unicellular red alga *Cyanidioschyzon merolae*. *DNA Res.* **10**: 67-77.

- Oliveira, M. C. and D. Bhattacharya (2000). Phylogeny of the Bangiophyticae (Rhodophyta) and the secondary endosymbiotic origin of algal plastids. *Am. J. Bot.* **87**: 482-492.
- Palmer, J. (2003). The symbiotic birth and spread of plastids: How many times and whodunit? *J Phycol.* **39**: 4-11.
- Palmer, J. D. and C. F. Delwiche (1998). The origin and evolution of plastids and their genomes. *Molecular Systematics of Plants II*. J. J. Doyle. Boston, Kluwer Academic Publishers pp. 375-409.
- Parke, M., I. Manton and B. Clarke (1955). Studies on marine flagellates II. Three new species of *Chrysochromulina*. *Journal of the marine biological association of the United Kingdom* **34**: 579-609.
- Pasternak, A., Y. Arashkevich and Y. Sorokin (1984). The role of the parasitic algal genus *Blastodinium* in the ecology of planktonic copepods. *Oceanology* **24**: 748-751.
- Patterson, D. J. (1999). The Diversity of Eukaryotes. *Am. Nat.* **154**: S96-S124.
- Patterson, D. J. (1989). Stramenopiles: chromophytes from a protistan perspective. *The chromophyte algae problems and perspectives*. J. C. Green, B. S. C. Leadbeater and W. L. Diver (eds). Oxford, Clarendon Press pp. 357-379.

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

Pfiester, L. and D. Anderson (1987). Dinoflagellate reproduction. *The Biology of Dinoflagellates*. F. J. R. Taylor (ed). Oxford, Blackwell Scientific.

Pfiester, L. and J. Popofsky (1979). Parasitic amoeboid dinoflagellates. *Nature* **279**: 421-424.

Piao, Y., N. T. Ko, M. K. Lim and M. S. H. Ko (2001). Construction of long-transcript enriched cDNA libraries from submicrogram amounts of total RNAs by a universal PCR amplification method. *Genome Research* **11**: 1553-1558.

Pierce, S. K., S. E. Massey, J. J. Hanten and N. E. Curtis (2003). Horizontal transfer of functional nuclear genes between multicellular organisms. *Biological Bulletin* **204**: 237-240.

Prezelin, B. (1987). Photosynthetic physiology of dinoflagellates. *The Biology of Dinoflagellates*. F. J. R. Taylor (ed). Oxford, Blackwell.

Race, H. L., R. G. Herrmann and W. Martin (1999). Why have organelles retained genomes? *TIG* **15**: 364-370.

- Rae, P. M. and R. E. Steele (1978). Modified bases in the DNAs of unicellular eukaryotes: An examination of distributions and possible roles, with emphasis on hydroxymethyluracil in dinoflagellates. *Biosystems* **10**: 37-53.
- Rambaut, A. and N. C. Grassly (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Applic. Biosci.* **13**: 235-238.
- Reece, K. S., M. E. Siddall, E. M. Burreson, and J. E. Graves (1997). Phylogenetic analysis of *Perkinsus* based on Actin Gene Sequences. *J. Parasitol.* **83**: 417-423.
- Reith, M. and J. Munholland (1995). Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. *Plant Molecular Biology Reporter* **13**: 333-335.
- Rhee, K. (2001). Photosystem II: The Solid Structural Era. *Ann. Rev. Plant Phys.* **30**: 307-328.
- Ris, H. and D. F. Kubai (1974). An unusual mitotic mechanism in the parasitic protozoan *Syndinium sp.* *J Cell Biol.* **60**: 702-720.
- Rizzo, P. J. (1987). Biochemistry of the dinoflagellate nucleus. *The Biology of Dinoflagellates*. F. J. R. Taylor (ed). Oxford, UK, Blackwell Scientific Publications: 143-173.

- Rizzo, P. J. and L. D. Noodén (1972). Chromosomal proteins in the dinoflagellate alga *Gyrodinium cohnii*. *Science*. **176**: 796-7.
- Rowan, R., S. M. Whitney, A. Fowler and D. Yellowlees (1996). Rubisco in marine symbiotic dinoflagellates: form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family. *The Plant Cell* **8**: 539-553.
- Sagan, L. (1967). On origin of mitosing cells. *J of Theoretical Biol* **14**: 225.
- Saldarriaga, J. F., F. J. Taylor, P. J. Keeling and T. Cavalier-Smith (2001). Dinoflagellate nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. *J. Mol. Evol.* **53**: 204-213.
- Saldarriaga, J. F., M. L. McEwan, N. M. Fast, F. J. R. Taylor and P. J. Keeling (2003). Multiple protein phylogenies show that *Oxyrrhis marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. *Int. J Syst Evol Microbiol* **53**: 355-365.
- Sambrook, J., E. F. Fritsch and T. Maniatis (1989). *Molecular cloning: a laboratory manual*. Plainview, Cold Spring Harbor Laboratory Press.
- Sánchez Puerta, M. V., T. R. Bachvaroff and C. F. Delwiche (2004). The complete mitochondrial genome of the haptophyte *Emiliana huxleyi* and its relation to heterokonts. *DNA Res* **11**:1-10.

- Sanchez-Puerta, M. V., T. R. Bachvaroff and C. F. Delwiche (2004). Complete chloroplast genome of the haptophyte *Emiliana huxleyi* (personal communication).
- Sanderson, M. J. and H. B. Shaffer (2002). Troubleshooting molecular phylogenetic analysis. *Annu. Rev. Ecol. Syst.* **33**: 49-72.
- Sarich, V. M. and A. C. Wilson (1973). Generation time and genomic evolution in primates. *Science.* **179**: 1144-1147.
- Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu and S. Tabata (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res* **6**: 283-290.
- Saunders, G. W., D. R. A. Hill, J. P. Sexton and R. A. Andersen (1997). Small-subunit ribosomal RNA sequences from selected dinoflagellates: testing classical evolutionary hypotheses with molecular systematic methods. *Origins of Algae and their Plastids*. D. Bhattacharya. New York, Springer Wein.
- Schein, A. I., J. C. Kissinger and L. H. Ungar (2001). Chloroplast transit peptide prediction: a peek inside the black box. *Nucleic Acids Research* **29**: art. no.-e82.
- Shalchian-Tabrizi, K. (2003). The evolution of dinoflagellates and their chloroplasts. Department of Biology. Oslo, University of Oslo.

- Shalchian-Tabrizi, K (1999). On the Evolution of Dinoflagellates. Masters Thesis, University of Oslo, Norway.
- Shimodaira, H. (2000). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**: 492-508.
- Shimodaira, H. and M. Hasegawa (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**: 1246-1247.
- Spector, D. L. (1984). Dinoflagellate Nuclei. *Dinoflagellates*. D. L. Spector (ed). London, Academic Press pp. 107-148.
- Steidinger, K. A. and K. Tangen (1997). *Dinoflagellates. Identifying marine phytoplankton*. C. R. Tomas. San Diego, California, Academic press.
- Stirewalt, V., C. Michalowski, W. Löffelhardt, H. Bohnert and D. Bryant (1995). Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*. *Plant Molecular Biology Reporter* **13**: 327-332.
- Strain, H., W. Svec, K. Aitzetmuller, M. Grandolfo, J. Katz, H. Kjosen, S. Norgard, S. Liaaen-Jensen, F. Haxo, P. Wegfarht and H. Rapoport (1971). The structure of peridinin, the characteristic dinoflagellate carotenoid. *J Am Chem Soc* **93**: 1823-1825.

Swift, I. E., and B. V. Milborrow (1981). Stereochemistry of allene biosynthesis and the formation of the acetylenic carotenoid diadinoxanthin and peridinin from neoxanthin. *Biochem. J.* **199**: 69-74.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis (1996). Phylogenetic Inference. *Molecular Systematics*. D. M. Hillis, C. Moritz and B. K. Mable (eds). Sunderland, Sinauer Associates, Inc.: 407-514.

Swofford, D. L., G. J. Olsen, P. J. Waddell, D. M. Hillis (2002). PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods). Version 4. Sunderland, Sinauer Associates.

Takishita, K., M. Ishikura, K. Koike and T. Maruyama (2003). Comparison of phylogenies based on nuclear-encoded SSUU rDNA and plastid-encoded *psbA* in the symbiotic dinoflagellate genus *Symbiodinium*. *Phycologia* **42**: 285-291.

Takishita, K., K. Koike, T. Maruyama and T. Ogata (2002). Molecular evidence for plastid robbery (Kleptoplastidy) in *Dinophysis*, a dinoflagellate causing diarrhetic shellfish poisoning. *Protist* **153**: 293-302.

Takishita, K. and A. Uchida (1999). Molecular cloning and nucleotide sequence analysis of *psbA* from the dinoflagellates: Origin of the dinoflagellate plastid. *Phycol. Res.* **47**: 207-16.

- Tappan, H. (1980). The Paleobiology of Plant Protists. San Francisco, W.H. Freeman and Co.
- Taylor, F. J. R. (1987). General group characteristics. The biology of dinoflagellates. F. J. R. Taylor (ed). Oxford, Blackwell.
- Taylor, F. J. R. Phylum Dinoflagellata. Handbook of Protoctista. L. Margulis, J.O. Corliss, M. Melkonian, and D.J. Chapman (eds), Jones and Bartlett, Boston.
- Tengs, T., O. J. Dahlberg, K. Shalchian-Tabrizi, D. Klaveness, K. Rudi, C. F. Delwiche and K. S. Jakobsen (2000). Phylogenetic analyses indicate that the 19<sup>th</sup> hexanoyloxy-fucoxanthin-containing dinoflagellates have tertiary plastids of haptophyte origin. Mol Biol Evol **17**: 718-29.
- Thomsen, H. A., K. R. Buck and F. P. Chavez (1994). Haptophytes as components of marine phytoplankton. The haptophyte algae. J. C. Green and B. S. C. Leadbeater (eds). Oxford, Clarendon Press. **51**: 187-208.
- Tomas, R. N. and E. R. Cox (1973). Observations on the symbiosis of *Peridinium balticum* and its intracellular alga. I. Ultrastructure. J Phycol **9**: 304-323.
- Turmel, M., C. Otis and C. Lemieux (1999). The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. Proc. Natl. Acad. Sci USA **96**: 10248-10253.

Turmel, M., C. Otis and C. Lemieux (2002). The chloroplast and mitochondrial genome of the charophyte *Chaetosphaeridium globosum*: insights into the timing of events that restructured organelle DNAs with the green algal lineage that led to land plants. *Proc. Natl. Acad. Sci USA* **99**: 11275-11280.

Uchida, A., S. Ebata, K. Wada, H. Matsubara and Y. Ishida (1988). Complete amino acid sequence of ferredoxin from *Peridinium bipes* (Dinophyceae). *J. Biochem* **104**: 700-705.

Van de Peer, Y. and R. De Wachter (1997). Evolutionary relationships among eukaryotic crown taxa taking into account site-to-site variation in 18S rRNA. *J. Mol. Evol.* **45**: 619-630.

Van de Peer, Y., S. A. Resing, U.-G. Maier and R. D. Wachter (1996). Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. *Proc. Natl. Acad. Sci USA* **93**: 7732-7736.

Van den Hoek, C., D. G. Mann and H. M. Jahns (1995). *Algae: an introduction to phycology*. New York, Cambridge University Press.

Vivier, E. and I. Desportes (1989). Phylum Apicomplexa. *Handbook of Protoctista*. L. Margulis, J. O. Corliss, M. Melkonian and D. J. Chapman (eds). Boston, MA, Jones and Bartlett: 549-573.

- Watanabe, M. M., S. Suda, I. Inouye, T. Sawaguchi and M. Chihara (1990).  
*Lepidodinium viride* gen et sp. nov. (Gymnodiniales, Dinophyta), a green  
dinoflagellate with a chlorophyll *a*- and *b*- containing endosymbiont. *J Phycol* **26**:  
741-751.
- Westbroek, P., J. Van Hinte, G. Brummer, M. Velhuis, C. Brownlee, J. Green, R. Harris  
and B. Heimdal (1994). The haptophyte algae. The haptophyte algae. J. C. Green  
and B. S. C. Leadbeater (eds). Oxford, Clarendon press.
- Whitney, S. M., D. C. Shaw and D. Yellowlees (1995). Evidence that some  
dinoflagellates contain a ribulose-1,5-bisphosphate carboxylase/oxygenase related  
to that of the alpha-proteobacteria. *Proc R Soc Lond B Biol Sci* **259**: 271-275.
- Wilson, R. J., and D. H. Williamson (1997). Extrachromosomal DNA in the  
Apicomplexa. *Microbiology and Molecular Biology Reviews* **61**: 1-16.
- Wolters, J. (1991). The troublesome parasites -- molecular and morphological evidence  
that Apicomplexa belong to the dinoflagellate-ciliate clade. *Biosystems* **25**:75-83.
- Yang, Z., R. Nelsen, N. Goldman and A. K. Pedersen (1997). PAML: a program package  
for phylogenetic analysis by maximum likelihood. *CABIOS* **13**:555-556.

- Yoon, H. S., J. D. Hackett and D. Bhattacharya (2002). A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis. *Proc. Natl. Acad. Sci USA* **99**: 15507-15512.
- Yoon, H. S., J. D. Hackett, G. Pinto and D. Bhattacharya (2002). The single, ancient origin of chromist plastids. *Proc. Natl. Acad. Sci USA* **99**: 15507-15512.
- Young, J. R., P. R. Bown and J. A. Burnett (1994). Paleontological perspectives. The haptophyte algae. J. C. Green and B. S. C. Leadbeater (eds). Oxford, Clarendon Press. **51**: 379-392.
- Zhang, Z. D., B. R. Green and T. Cavalier-Smith (1999). Single gene circles in dinoflagellate chloroplast genomes. *Nature* **400**: 155-159.
- Zhang, Z. D., B. R. Green and T. Cavalier-Smith (2000). Phylogeny of ultra-rapidly evolving dinoflagellate chloroplast genes: A possible common origin for sporozoan and dinoflagellate plastids. *J. Mol. Evol.* **51**: 26-40.
- Zuegge, J., S. Ralph, M. Schmucker, G. I. McFadden and G. Schneider (2001). Deciphering apicoplast targeting signals -- feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* **280**: 19-26.