MOLECULAR EVOLUTIONARY STUDIES ON THE CHLORARACHNIOPHYTE *Bigelowiella natans*

by

MATTHEW BRIAN ROGERS

B.Sc., The University of Alberta, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

December 2006

© Matthew Brian Rogers, 2006
Abstract

Chlorarachniophytes are marine cercozoan amoeboflagellates with plastids derived from a secondary endosymbiotic event involving a green alga. The retention of a vestigial eukaryotic nucleus, or 'nucleomorph' in the plastid of chlorarachniophytes makes chlorarachniophytes ideal organisms for the study of secondary endosymbiosis. Among chlorarachniophytes, the majority of sequence data are from a single species, *Bigelowiella natans*. Thousands of expressed sequence tags and a complete nucleomorph and chloroplast genome from *Bigelowiella natans* provide a detailed picture of genes encoding proteins with function in the plastid of chlorarachniophytes. The phylogeny of three plastid-targeted Calvin cycle enzymes with complicated distributions and multiple recompatmentlalisation events are described. These three enzymes are also the only known Calvin cycle enzymes in *B. natans* that show any evidence of having been acquired through endosymbiotic gene transfer. Previous studies have detailed the contribution of lateral gene transfer to the plastid function of *B. natans*. Several specific examples of these events are expanded on and discussed. Plastid-targeting peptides from published sequences are characterized and a heterologous targeting experiment in to the apicoplast of *Toxoplasma gondii* is described. The complete plastid genome of *Bigelowiella natans* has been sequenced. This plastid genome appears to be the smallest among all known photosynthetic plastid genomes, though it nevertheless retains most of the photosynthesis related genes present in chlorophytes. Phylogenetic analysis of concatenated chloroplast protein coding genes indicates that the endosymbiont of *B. natans* may be a derived green algae, and argues strongly against a single origin of chlorarachniophyte and euglenids plastids predicted by the cabozoa hypothesis.
Table of Contents

Abstract ......................................................................................................................... ii

Table of Contents .......................................................................................................... iii

List of Tables ................................................................................................................... vii

List of Figures ................................................................................................................. viii

Acknowledgements ....................................................................................................... ix

Dedication ....................................................................................................................... x

Co-Authorship Statement ............................................................................................. xi

Chapter 1 - Introduction ............................................................................................ 1

1.1 Overview ................................................................................................................ 1

1.2 History and Landmarks ........................................................................................ 1

1.2.1 Discovery and Description ............................................................................. 1

1.2.2 A New Group of Eukaryotic Algae ................................................................. 2

1.3 The Nucleomorph genome of Bigelowiella natans ............................................. 5

1.3.1 The smallest known introns .......................................................................... 6

1.3.2 The “Raison D’Etre” of the Nucleomorph ....................................................... 7

1.3.3 Comparison to the Guillardia theta nucleomorph ........................................... 8

1.4 Controversy Over Endosymbiont Origins ............................................................ 9

1.5 Host affinities ......................................................................................................... 10

1.6 The Cabozoans ...................................................................................................... 11

1.7 The Chromalveolates ........................................................................................... 13

1.8 Structure and objectives of thesis ....................................................................... 14
Chapter 2 - Lateral Transfer and Recompartmentalisation of Calvin cycle enzymes of plants and algae ............................................................... 24

2.1 Introduction ............................................................................... 24

2.2 Methods .................................................................................. 26

2.2.1 Identification, sequencing, and primary analysis of FBA and FBPase/SBPase genes ................................................................. 26

2.3 Results and Discussion ............................................................. 28

2.3.1 Class I FBA ........................................................................ 28

2.3.2 Class II FBA ........................................................................ 30

2.3.3 FBPase/SBPase .................................................................... 34

2.4 Concluding remarks: the evolution of FBA, FBPase and SBPase. ................................................................. 37

2.5 Bibliography ............................................................................ 38

Chapter 3 - A complex and punctate distribution of three genes derived by lateral gene transfer in eukaryotic genomes ................................................................. 42

3.1 Introduction ............................................................................... 42

3.2 Methods .................................................................................. 45

3.2.1 Characterisation of new sequences ........................................ 45

3.2.2 Sequence analyses ............................................................... 46

3.3 Results and Discussion ............................................................. 47

3.3.1 Ribulose-5-phosphate-3-epimerase ........................................ 47

3.3.2 Glyceraldehyde-3-Phosphate Dehydrogenase .......................... 51

3.3.3 Transketolase ..................................................................... 55

3.4 Concluding Remarks ............................................................... 60

3.5 Bibliography ............................................................................ 62
Chapter 4 - Plastid-targeting peptides from the chlorarachniophyte *Bigelowiella natans* ........................................ 67

4.1 Introduction ............................................................................................................. 67

4.2 Materials and Methods .......................................................................................... 69

4.2.1 Assembling a data-set of cytosolic, ER, and plastid proteins from *B. natans*. .......................................................... 69

4.2.2 Signal peptide predictions and analysis ................................................................. 72

4.2.3 Transit peptide prediction and analysis ................................................................. 72

4.2.4 Heterologous activity of *B. natans* plastid-targeting leader in *T. gondii* .......... 73

4.3 Results and Discussion ......................................................................................... 74

4.3.1 *B. natans* Signal Peptides ................................................................................. 74

4.3.2 *B. natans* transit peptides ................................................................................. 77

4.3.3 Heterologous targeting in *T. gondii*. ................................................................. 81

4.4 Bibliography ........................................................................................................... 84

Chapter 5 - The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. ............... 88

5.1 Introduction ............................................................................................................. 88

5.2 Materials and Methods ......................................................................................... 90

5.2.1 Genome sequencing and annotation ................................................................. 90

5.2.2 Pulsed field gels ............................................................................................... 91

5.2.3 Phylogenetic analyses ....................................................................................... 91

5.3 Results and Discussion ......................................................................................... 93

5.3.1 Genome structure ............................................................................................. 93

5.3.2 Gene content, loss, and compaction ................................................................. 97

5.3.3 tRNA genes ..................................................................................................... 101
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3.4 Phylogenetic relationship to other plastid genomes</td>
<td>101</td>
</tr>
<tr>
<td>5.3.5 Origin of chlorarachniophyte plastids</td>
<td>105</td>
</tr>
<tr>
<td>5.4 Conclusions</td>
<td>106</td>
</tr>
<tr>
<td>5.5 Bibliography</td>
<td>108</td>
</tr>
<tr>
<td><strong>Chapter 6 - Conclusions</strong></td>
<td>114</td>
</tr>
<tr>
<td>6.1 Summary</td>
<td>114</td>
</tr>
<tr>
<td>6.2 Transfer and recompartmentalisation of carbon metabolic genes</td>
<td>114</td>
</tr>
<tr>
<td>6.3 Evidence for inter and intra domain transfer</td>
<td>116</td>
</tr>
<tr>
<td>6.4 Plastid-targeting leaders</td>
<td>118</td>
</tr>
<tr>
<td>6.5 The chloroplast genome of <em>Bigelowiella natans</em></td>
<td>119</td>
</tr>
<tr>
<td>6.6 General Conclusions</td>
<td>121</td>
</tr>
<tr>
<td>6.7 Bibliography</td>
<td>123</td>
</tr>
</tbody>
</table>
List of Tables

Table 4.1 New cytosolic, nuclear and endomembrane-targeted proteins in *Bigeloviella natans* ................................................................. 71
List of Figures

Figure 2.1 Protein maximum likelihood tree of class I FBA ........................................... 29
Figure 2.2 Composite protein maximum likelihood tree of type A and type B class II
    FBA .................................................................................................................. 32
Figure 2.3 Protein maximum likelihood phylogeny of FBPase and SBPase .................. 36
Figure 3.1 Bayesian tree of RPE with PROML branch lengths ................................ 48
Figure 3.2 Bayesian tree of GAPDH with PROML branch lengths .............................. 53
Figure 3.3 Bayesian tree of transketolase with PROML branch lengths ..................... 57
Figure 3.4 Boxshade alignment of a 4 amino acid indel common to diatoms, chlorarachniophytes and a dinoflagellate ............................................................... 58
Figure 4.1 Chemical properties of the signal/transit peptide boundary ...................... 76
Figure 4.2 Comparison of amino acid composition between transit peptides and mature
    and cytosolic peptides ...................................................................................... 78
Figure 4.3 Chemical properties of transit/mature peptide boundary ........................... 80
Figure 4.4 Heterologous expression of the *Bigelowiella natans* RuBisCo small subunit
    leader peptide in the apicomplexan *Toxoplasma gondii* .................................. 83
Figure 5.1 Chloroplast genome of *Bigelowiella natans* .......................................... 94
Figure 5.2 Pulse field gel electrophoresis of chloroplast genome chromosome ........ 95
Figure 5.3 Histogram of plastid genome sizes ......................................................... 98
Figure 5.4 Plastid gene loss events plotted on the phylogeny of concatenated plastid
    genes ............................................................................................................. 100
Figure 5.5 Protein maximum likelihood tree of concatenated plastid-encoded genes.... 103
Figure 5.6 Protein maximum likelihood tree of concatenated plastid-encoded genes.... 104
I would like to acknowledge the following individuals for their contribution to my work, both through direct contribution of results as well as useful guidance and support. J.M. Archibald for two important collaborations, as well as many important discussions on chlorarachniophyte and nucleomorph evolution. I would like to acknowledge P.R. Gilson, V. Su and G.I. Mcfadden for their collaboration on the B. natans chloroplast genome. I would like to acknowledge R.A. Waller for useful discussions on apicoplasts, organellar genomes and plastid-targeting. I would like to acknowledge M. Field for his assistance with software applications. I wish to thank A.P. deKoning for helpful advice and discussions on the subject of plastid genome evolution. I would like to acknowledge N.J. Patron for two collaborations and interesting discussions on nucleomorph and plastid evolution. I would like to thank J.T. Harper for advice on molecular biological techniques and his continuing friendship.

I would like to thank my supervisory committee, Drs. J. Bohlmann, B.R. Green and B.S. Leander for their suggestions and input concerning my research and thesis.

Finally and foremost I would like to thank P.J. Keeling for his support and guidance as my PhD supervisor.
Dedication

For Heather.
Co-Authorship Statement

The manuscript included as chapter 2 was originally published as: Rogers, M.B. and Keeling, P.J. 2004. Lateral transfer and recompartmentalisation of Calvin cycle enzymes of plants and algae. J. Mol. Evol. Apr. 58(4):367-75. cDNA library construction and EST sequencing were performed by P.J. Keeling. P.J. Keeling and myself contributed to the concept of this manuscript. I performed phylogenetic analyses under the supervision of P.J. Keeling. Manuscript was written by myself and revised by P.J. Keeling.

The manuscript included as chapter 3 is unpublished. Both myself and P.J. Keeling contributed to the concept of this chapter. EST projects of *Euglena gracilis* and *Hartmanella vermiciformis* and *Physarum polycephalum* were constructed by M.W. Gray and R. Watkins and D.G. Durnford. ESTs for each of these were provided by R.F. Watkins and re-sequenced by myself. EST projects of *I. galbana* and *K. micrum* were constructed by R.A. Waller and N.J. Patron. ESTs from these projects were re-sequenced by myself. EST project of *B. natans* was constructed by P.J. Keeling. Degenerate PCR product of *B. natans* transketolase was amplified, cloned and sequenced by myself. Degenerate PCR product of *I. galbana* GAPDH was amplified, cloned and sequenced by J.T. Harper. Data analysis was carried out by myself. Manuscript was written by myself and P.J. Keeling.

The manuscript included as chapter 4 was originally published as: Rogers M.B., Archibald J.M., Field M.A., Li C, Striepen B., Keeling P.J.. 2004. Plastid-targeting peptides from the chlorarachniophyte Bigelowiella natans. J. Eukaryot. Microbiol. Sep-Oct;51(5):529-35. EST project of *B. natans* was constructed by P.J. Keeling. Analysed plastid-targeted proteins were reported in an earlier manuscript by J.M. Archibald. Heterologous targeting of RubisCO in *Toxoplasma gondii* was performed by C. Li and B. Striepen. M. Field designed software for analysis of amino acid content and hydrophobicity of signal peptides and transit peptides. Analysis of signal and transit
peptides was carried out by myself. Manuscript was written by myself and revised by P.J. Keeling.

The manuscript included as chapter 5 was originally published Rogers, MB, Gilson, PR, Su, V, McFadden GI, Keeling PJ. 2006. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: Evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. Mol. Biol. Evol Sep 21 Epub ahead of print. Chloroplast genome sequence data were generated in large part through the *B. natans* nucleomorph sequencing project by P.R. Gilson and V. Su under the supervision of G.I. McFadden. Remaining chloroplast genome sequence were generated by myself and P.J. Keeling. Assembly, annotation and analysis of the chloroplast genome data was conducted by myself under the supervision of P.J. Keeling. Manuscript was written by myself and revised and edited by P.J. Keeling.
Chapter 1 - Introduction

1.1 Overview

Chlorarachniophytes are marine cercozoan amoebae or flagellates with chloroplasts derived from secondary endosymbiosis. Chlorarachniophytes are among the least speciose groups of algae, and are often ignored among the greater diversity of algae. Nevertheless, this group occupies a position of crucial importance with respect to the theory of secondary endosymbiosis. As one of only two groups of algae where the nucleus of the engulfed algal endosymbiont remains as a reduced "nucleomorph", chlorarachniophytes are powerful evidence in support of the theory of secondary endosymbiosis. Below is a brief summary of the history of discovery of chlorarachniophytes and characterization of the chlorarachniophyte nucleomorph.

1.2 History and Landmarks

1.2.1 Discovery and Description

By the standards of most groups of algae, chlorarachniophytes are a very recently discovered group, having only been described 20 years ago. (Hibberd and Norris 1984) The type specimen for this group Chlorarachnion reptans was first cultured from the Canary Islands (Hibberd and Norris 1984) by L. Geitler in 1930 who described a photosynthetic filose amoeboid cell with both an amoeboid stage as well as a walled-cyst stage. C. reptans was later discovered to have a flagellated stage as well. Amoeboid, walled cyst and flagellate zoospores are morphologies found in many chlorarachniophytes, often with one morphology existing as the primary vegetative stage, and occasionally one morphology being completely absent or as yet unobserved. For example, the genus Gymnochlora stellata has an amoeboid vegetative stage but also forms coccoid zoospores; a flagellate stage has never been observed (Ishida, Nakayama, and Hara 1996). Geitler named his discovery Chlorarachnion reptans, describing the reticulating spider-web like networks, or plasmodia that these cells formed from fused filopodia (Moestrup and Sengco 2001). Geitler classified C. reptans among the heterokonts based on its green pigmentation, its plasmodial morphology (observed in
some heterokonts) and the absence of starch or paramylon storage products, but nevertheless considered this classification to be only provisional (Hibberd and Norris 1984).

1.2.2 A New Group of Eukaryotic Algae

Hibberd and Norris (Hibberd and Norris 1984) re-classified *C. reptans* among a unique group of eukaryotes on the basis of light and electron microscope observations and pigment composition. Hibberd and Norris (1984) described cells with filose pseudopodia linked together to form plasmodial webs, and between five and seven chloroplasts with prominent pyrenoids encircling a central nucleus. Hibberd and Norris (1984) also found a flagellate stage of *C. reptans*. Pigment analyses revealed that chlorophyll *a* and *b* were present but chlorophyll *c* and phycobilins were absent. As chlorophyll C was characteristic of red algae and chromistan algae (algae with plastids derived from secondary endosymbiosis with red algae) and chlorophyll b was found solely in green algae, plants and euglenids, *C. reptans* clearly had more in common with the latter groups (Hibberd and Norris 1984). Nevertheless, *C. reptans* was ultrastructurally unlike any chlorophyte or euglenid in having an amoeboid morphology.

The plastids of *C. reptans* were observed to have four membranes surrounding them, more than those found in green algae where only two were present, and more than in euglenid and dinoflagellate algae where three were known to be present. The presence of multiple membranes surrounding a plastid were at this time known to be consistent with a secondary endosymbiotic event, as had been previously argued for the plastids of euglenids (Gibbs 1978), dinoflagellates (Gibbs 1981b), cryptophytes (Greenwood, Griffiths, and Santore 1977) and related chromistans (Gibbs 1981b). The plastid ultrastructure of *C. reptans* was in this regard suggestive of a eukaryotic endosymbiont incorporated into an amoeboid cell. Like chlorarachniophytes, chromistans, a group consisting of heterokonts, haptophytes and cryptomonads, have four membranes surrounding their plastid. However, the plastids of chromistans are unlike those of chlorarachniophytes in that the outer membrane often consists of rough ER, termed chloroplast ER or CER (Bouck 1965; Gibbs 1970; Hibberd and Leedale 1970). Although Hibberd and Norris (1984) determined that the outer membrane of the chloroplast was
distinct from the rough ER of chromistan plastids and no connections to ER elsewhere in the cell were observed, they nevertheless referred to the outer membrane as CER in the absence of a better term. Hibberd and Norris (1984) also noted a distinct compartment at the base of the pyrenoid, enclosed between the inner two and outer two membranes of the plastid. A similar compartment had previously been described from cryptophytes, and termed the periplastidial space in recognition of its position between two membranes making up the plastid envelope, and the two outer host derived membranes of the endosymbiont (Gibbs 1981a). In addition to ribosome-like particles, Hibberd and Norris (1984) also observed a double membrane bound organelle in the periplastidial space, embedded in a groove in the pyrenoid of the chloroplast. Similar structures had been observed in members of the cryptophyceae, where they too were often situated in a groove in the pyrenoid (Gillott and Gibbs 1980). This structure had been interpreted by various investigators to be the remains of a eukaryotic nucleus in an endosymbiotic organelle derived through secondary endosymbiosis (Greenwood, Griffiths, and Santore 1977; Whatley 1981; Morrall and Greenwood 1982). Greenwood coined the term 'nucleomorph' to describe this tiny eukaryotic nucleus (Greenwood 1974). Although not identical to the nucleomorph of cryptophytes, Hibberd and Norris (1984) predicted (correctly) that this organelle was in fact a nucleomorph similar to that of cryptophytes, though they had no evidence that nucleic acids were in fact contained within this organelle. In the end, Hibberd and Norris (1984) concluded that *C. reptans* was derived from an endosymbiosis between a heterotrophic amoeboid cell and a phototrophic eukaryotic. The plastids of *C. reptans* were derived from a chlorophyll b containing eukaryote, and in a manner similar to that described in other algae with plastids with greater than two membranes, this alga had been reduced to a photosynthetic organelle. The phylogenetic relationship of the host organism to other groups of eukaryotes could not be determined. *C. reptans* was unlike any other group of green phototrophic eukaryotes in its morphology.

Treatment of the chloroplasts of *C. reptans* with DAPI stain showed that the double membrane bound organelle in the periplastid space did indeed contain nucleic acids (Ludwig and Gibbs 1989). In addition, these nucleic acids proved to be vulnerable to DNAse, but not RNAse digestion, demonstrating conclusively that this membrane
bound structure was in fact a nucleomorph. Similar studies had shown the same was true for cryptophyte nucleomorphs (Gillott and Gibbs 1980; Ludwig and Gibbs 1985) and had demonstrated that the nucleomorph of cryptophytes contained a prominent nucleolus like structure, a feature that verified the eukaryotic nature of this organelle.

The chloroplasts of chlorarachniophytes are in some ways similar to those of euglenids in having prominent pyrenoids and cytosolically stored paramylon. On this basis, Cavalier-Smith (Cavalier-Smith 1986) proposed that the endosymbiont might have been a euglenid. Ludwig and Gibbs (1989) argued against this hypothesis, as three membranes surround euglenid plastids. A euglenid-derived plastid would be surrounded by five membranes rather than four. Even if the phagocytic vesicle surrounding the engulfed euglenid had disappeared, the position of the nucleomorph between the chloroplast envelope membranes and the outer two membranes could not be explained by this hypothesis. If a euglenid had been the symbiont and the phagocytic vesicle had been lost subsequent to uptake of this cell, then the nucleomorph would be located between the outer (plasma) membrane and the inner three membranes (Ludwig and Gibbs 1989). Like Hibberd and Norris, Ludwig and Gibbs argued for a chlorophycean origin for the plastid of *C. reptans*. These predictions were soon validated by molecular data.

Although there was strong microscopic evidence that the nucleomorph of *C. reptans* was a miniature eukaryotic nucleus, little was known about the composition of this genome. The first gene to be sequenced from a nucleomorph genome was the small subunit ribosomal RNA from the cryptophyte *Guillardia theta* (Douglas et al. 1991). Soon after, nuclear and nucleomorph SSU rRNA was sequenced from four different chlorarachniophytes (McFadden et al. 1994). In-situ hybridization of these clones from *Chlorarachnion* CCMP 242 showed one copy localized to the host nucleus, the other localized to the nucleolus of the nucleomorph and ribosomes in the periplastid space (McFadden et al. 1994). Specific probes for nucleomorph rRNA genes of *Chlorarachnion* CCMP 242 localized a copy of this gene to three tiny chromosomes 145 KB, 140 kbps and 95 kbps in size (McFadden et al. 1994). The possibility remained of course, that other nucleomorph chromosomes not encoding rRNAs could be found. This possibility was later precluded by southern blot analysis using telomeric clones in the related strain *Bigelowiella natans*. The nucleomorph genome of *Chlorarachnion* CCMP
242 was thus found to be 380 kbps in size with a karyotype of 3 chromosomes. This size and karyotype is very similar to that of *Bigelowiella natans*, the species in which the nucleomorph genome has been sequenced.

1.3 The Nucleomorph genome of *Bigelowiella natans*

Although preliminary characterization of the chlorarachniophyte nucleomorph was in *Chlorarachnion* CCMP 242, *Bigelowiella natans* was selected for a nucleomorph genome project, as it grows to high density in aerated cultures (Gilson and McFadden 2002). The nucleomorph genome of *Bigelowiella natans* is to date completely sequenced (Gilson et al. 2006). Sequencing of the *B. natans* nucleomorph was preceded by the sequencing of the telomeric regions of chromosome III (Gilson and McFadden 1996) and sequencing of a 13.2 kbp restriction fragment of nucleomorph chromosome III (Gilson and McFadden 1995). Telomeric regions had been identified by restriction digests of chromosome III, and selection of cloned fragments containing a restriction digested sticky-end and an endonuclease-resistant blunt-end presumably containing a telomere. Sequencing of a clone with an endonuclease-resistant end revealed telomeres consisting of a 7 base-pair motif (TCTAGGG) repeated 32 times (Gilson and McFadden 1995). Sequencing the opposite end of the clone revealed the presence of a gene encoding 5.8S rRNA, and the previously described 18S rRNA gene, confirming the telomere did belong to the nucleomorph. Southern-blot analysis of restriction digested fragments of whole cell DNA using the telomere sequence as a probe revealed variable sizes of telomeres across restriction fragments. Gilson and McFadden (1995) concluded that this signified that the number of telomere repeats was variable across chromosomes as the number of telomeric repeats is known to vary within the life-time of a chromosome. They further demonstrated that the same probe hybridized to pulse-field gel blots of all three nucleomorph chromosomes, arguing strongly for a 3 chromosome karyotype for the nucleomorph of *Bigelowiella natans* (Gilson and McFadden 1995).

The discovery of further nucleomorph genes came from the sequencing of a 13.2 kbp fragment previously digested from nucleomorph chromosome III and the creation of a cDNA library from total *B. natans* RNA. From the 13.2 kb fragment, 2 ribosomal proteins were identified, as well as components of a spliceosome, subunits of clpP
protease, a RNA polymerase subunit and part of an RNA helicase (Gilson and McFadden 1996). Expression of the spliceosomal protein was verified through in-situ hybridization and localized to the nucleomorph. Antibodies constructed against subunits of various spliceosomal subunits localized spliceosomal subunits to both the nucleomorph and host nucleus (Gilson and McFadden 1996).

1.3.1 The smallest known introns

The nucleomorph of *Bigelowiella natans* encodes the smallest known spliceosomal introns. Three sizes of introns were originally described: 18, 19 and 20 base-pair introns (Gilson and McFadden 1996). With the completion of the genome, this range was expanded to include 21 base pair introns (Gilson et al. 2006). Regardless of the upper limits on nucleomorph intron sizes, 18 bps is the smallest eukaryotic intron found to date, though it is rivaled by introns in *Paramecium tetraurelia*, which are often as small as 20 bps (Dupuis 1992; Russell, Fraga, and Hinrichsen 1994). The introns of chlorarachniophytes have canonical splice sites (5' GT...AG 3') like most eukaryotic spliceosomal introns. No internal consensus sequence could be detected, though the intron sequences displayed a tendency towards high AT content, which may serve as a basis for detection by the spliceosome in conjunction with the introns' boundaries (Gilson et al. 2006). At first glance, the nucleomorph of *B. natans* contains an immense number of introns, 852 introns were identified in the completed nucleomorph genome (Gilson et al. 2006). In comparison, the nucleomorph of the cryptophyte *Guillardia theta* contains only 17 introns. Only one intron with a homologous position in another eukaryote was identified initially (Gilson and McFadden 1996), but with the completion of the nucleomorph genome, and the *Chlamydomonas reinhardtii* nuclear genome a larger scale comparison became possible. Gilson et al. (2006), compared 137 introns from 44 conserved nucleomorph genes with introns in *Chlamydomonas* and *Arabidopsis* and found that 77% had homologous positions in either species, and 38% were shared with both. These findings suggested both that the introns present in the nucleomorph genome were in large part present ancestrally, and that the intron density of the *B. natans* nucleomorph was similar to that of plants and green algae. Although it was clear that the
introns of *B. natans* had suffered massive reductions in size, the ancestral intron density of the nucleomorph seems to have been conserved.

1.3.2 The "Raison D’Etre" of the Nucleomorph

Gilson and McFadden originally speculated that the "raison-d’etre" of the nucleomorph was encoding proteins necessary for plastid function (Gilson and McFadden 1996). Although the bulk of the genes encoded by the nucleomorph exist for the expression and replication of the nucleomorph itself, the existence of just one essential chloroplast gene could explain why nucleomorphs persist in some while those of so many other lineages of secondary-symbiont containing eukaryotes have gone extinct (Gilson and McFadden 1996). A possible candidate was identified early on; a gene encoding a subunit of clpP protease with an N-terminal extension (compared to plastid-encoded versions of the same gene) was identified from the 13.2-kbp restriction fragment of nucleomorph chromosome III (Gilson and McFadden 1996). With the completion of the nucleomorph genome, 17 other putative plastid genes have been identified, including many more subunits of clpP protease, as well as plastid sigma factors, the inner membrane import protein tic20, chaperonins and others (Gilson et al. 2006). By comparison, the completed nucleomorph of the cryptophyte *Guillardia theta* encodes 30 plastid proteins (Douglas et al. 1991), a few of which are related to plastid targeted proteins in *B. natans* (see below) (Gilson et al. 2006). With the completion of genomes from other organisms with secondary endosymbionts lacking nucleomorphs, such as the diatom *Thalassiosira pseudonana* (Armbrust et al. 2004) and the apicomplexan *Plasmodium falciparum* (Gardner et al. 2002), it is apparent that many plastid-targeted nucleomorph encoded genes could, and have been transferred to the host nucleus in the course of nucleomorph reduction in these lineages. In the end, Gilson et al. (2006) conclude that no single nucleomorph gene could account for the existence of the structure as a whole (Gilson et al. 2006). The retention of a nucleomorph might best be viewed as a historical accident. Whereas loss of nucleomorphs has occurred at least twice (see below), the retention of nucleomorphs might simply be the result of endosymbiotic gene transfer events that have yet to occur.
1.3.3 Comparison to the Guillardia theta nucleomorph

Striking similarities and differences exist between the two completely sequenced nucleomorph genomes. Superficially, both nucleomorphs are similar in that they have the same karyotype (3 chromosomes), and are roughly similar in total size (373 kbps in B. natans vs. 551 kbps in G. theta) in addition to chromosome size (100-200 kbps per chromosome). Many similarities in gene content exist between the two genomes, and housekeeping genes are over-represented in both (Gilson et al. 2006). DNA modifying enzymes and DNA polymerases are absent from both and the same aminoacyl tRNA synthetase, serine aminoacyl tRNA synthetase is present in both, though all other aminoacyl tRNA synthetases are curiously absent in both (Gilson et al. 2006). Both have chromosomes capped with rDNA genes linked to telomeres, though the telomeres of the two are clearly distinct in sequence and the rDNA genes are transcribed in opposite directions (Gilson and McFadden 1995; Zauner et al. 2000). This curious similarity has been linked to the need for high sequence fidelity in rRNA genes. Subtelomeric regions undergo extensive recombination, thus the presence of essential rRNA genes in this region might be linked to the maintenance of sequence conservation in these genes through gene conversion events (Gilson and McFadden 2002). Both share similarities related to massive genome reduction, including small intergenic spaces, a high coding to non-coding sequence ratio and transcripts containing coding sequence from upstream and downstream genes due likely to the use of upstream and downstream genes as transcriptional promoters and terminators (Douglas et al. 2001; Gilson and McFadden 2002; Williams et al. 2005; Gilson et al. 2006). In spite of these similarities, the two nucleomorphs differ in terms of gene content. The G. theta nucleomorph differs in that it encodes tubulin subunits, proteasome subunits, 5S rRNA, telomerase RNA, U4 small RNA and tRNAs not present in the B. natans nucleomorph (Gilson et al. 2006). As described above, 30 plastid proteins are encoded in the G. theta nucleomorph, whereas 17 are encoded in the B. natans nucleomorph. Only three of these genes are shared by the two nucleomorphs, two encoding subunits of clpP protease, the other encoding GroEl. Of these three, the two GroEls are not orthologues, and the complexity of the clpP family
makes it difficult to infer homology between the two proposed clpP homologues (Gilson et al. 2006). Unlike the "pygmy" introns of B. natans, the introns of G. theta are within the standard range of introns sizes (42-52 bps) and there are far fewer of them. Only 17 introns are present in the G. theta nucleomorph, compared to the 852 in the B. natans nucleomorph. This may relate to the observation that known red algal genomes are intron-poor compared to green algal and plant genomes (Matsuzaki et al. 2004; Gilson et al. 2006). With the completion of the B. natans genome project, it is clear that cryptophyte and chlorarachniophyte nucleomorphs represent independent, but surprisingly convergent experiments in genome reduction.

1.4 Controversy Over Endosymbiont Origins

The symbiont of chlorarachniophytes has long been recognized as a green alga based on the presence of green algal pigments such as chlorophyll b although few ultrastructural characters have supported this relationship. Discovery and sequencing of SSU rRNA from the nucleomorph and cytosol of Bigelowiella natans provided irrefutable proof of the eukaryotic nature of the chlorarachniophyte plastid but these sequences showed little affinity for any known eukaryotic group (McFadden et al. 1994). The paucity of available eukaryotic sequences as well as the rapid rate of nucleomorph evolution and suggested antiquity of the chlorarachniophyte partnership (McFadden et al. 1994) accounted for the ambiguous relationship of these sequences. Previous studies using nucleomorph and nuclear 18S rRNA had suggested a possible relationship between the nucleomorphs of chlorarachniophytes and that of the cryptophyte Guillardia theta (Cavalier-Smith 1993; Cavalier-Smith, Allsopp, and Chao 1994), suggesting a unique origin of all nucleomorph containing lineages. The first plastid encoded proteins to be sequenced from a chlorarachniophyte, small subunit ribosomal RNA and large subunit RubisCo placed chlorarachniophytes at the base of a group consisting of chlorophytes, euglenids and plants when corrected for substitutional biases in AT content using LogDet transformation (McFadden, Gilson, and Waller 1995). With new methods in substitution rate calibration, Van de Peer and his colleagues (Van de Peer, Rensing, and Maier 1996) constructed phylogenies of nucleomorph 18S rRNA that placed chlorarachniophytes within the chlorophyte with weak bootstrap support. Removing Guillardia theta
nucleomorph 18S rRNA from the analysis greatly augmented the support for this relationship. Van de Peer speculated that a degree of long branch attraction was occurring between the two rapidly evolving nucleomorph sequences, causing them to artefactually branch together in previous 18S rRNA phylogenies (Van de Peer, Rensing, and Maier 1996).

With the endosymbiont of chlorarachniophytes now known to be of chlorophyte origin, the precise relationships of the chlorarachniophyte endosymbiont to extant groups of green algae could be addressed. Van De Peer's (1996) phylogeny supported a weak relationship to trebouxiophyceans. Rare-pigment data suggested a relationship to prasinophycean green algae (Sasa et al. 1992). Phylogenetic analysis of plastid-encoded tufA genes produced trees with strong bootstrap support for a sister relationship between the chlorarachniophyte Gymnochlora stellata and ulvophyte green algae (Ishida et al. 1997). This relationship was further supported by nucleomorph SSU rRNA phylogeny including a broader range of chlorarachniophyte and green algal sequences (Ishida, Green, and Cavalier-Smith 1999). Though molecular evidence thus far seems to support an ulvophycean origin, few developments have been made since in determining the affinities of the chlorarachniophyte symbiont. Recent phylogenies of host-encoded plastid-targeted genes revealed many genes with green algal affinities in addition to a large number of what appear to be laterally transferred genes from eubacteria and other eukaryotes. Unfortunately, the paucity of available green algal nuclear data prevented any specific determination of the nature of the chlorarachniophyte endosymbiont in trees that did suggest green algal ancestry (Archibald et al. 2003).

1.5 Host affinities

The phylogenetic relationship of the host component of chlorarachniophytes is now well understood, though it was a subject of past contention. The complicated life cycle and cellular ultrastructure of chlorarachniophytes do not suggest a close relationship with any existing group of eukaryotes (Hibberd and Norris 1984). As a result of this ambiguity, chlorarachniophytes have been summarily classified among the heterokonts, euglenids and cryptophytes (Geitler 1930; Cavalier-Smith 1986; Grell 1990; Cavalier-Smith, Allsopp, and Chao 1994). In recent years, molecular phylogenetics has
placed the host lineage of chlorarachniophytes among the cercozoa. Cercozoa are a
diverse group of eukaryotes in morphology and habitat. Cercozoa include such varied
forms as plasmodium-forming plant and animal parasites, thecate filose amoebae, soil
amoeboflagellates and organisms superficially similar to radiolarians. (Cavalier-Smith
1998; Polet et al. 2004). Current eukaryotic classifications place cercozoa within a larger
group of eukaryotes, including radiolarians and foraminiferans collectively referred to as
rhizarians.

Small subunit rRNA phylogeny initially suggested a relationship between
chlorarachniophytes and euglyphids (Bhattacharya, Helmchen, and Melkonian 1995).
Subsequent phylogenies of SSU rRNA including plasmodiophorids and other groups of
filose amoebae succeeded in recovering this same relationship with *Chlorarachnion
reptans* branching as sister to a polyphyletic group of euglyphids and other filose
amoebae but within a larger group including plasmodiophorids and filose amoebae
(Cavalier-Smith 1996/1997). The first nuclear-encoded protein coding genes amplified
from a chlorarachniophyte, alpha and beta-tubulin provided further support for the host of
chlorarachniophytes being a relative of filose amoebae (Keeling, Deane, and McFadden
1998). Actin-encoding genes from *Bigelowiella natans* and *Lothorella amoebiformis*
recovered a clade consisting of chlorarachniophytes, cercozoans and foraminiferans,
grouping cercozoa with foraminiferans for the first time (Keeling 2001). More recently,
ubiquitin sequences of *Bigelowiella natans* have been shown to contain indels unique to
rhizarians, providing solid gene character support for the chlorarachniophyte host as a
member of the rhizaria (Archibald et al. 2002). Recent phylogenies including a wider
sampling of cercozoans suggest that the chlorarachniophyte host lineage is a cercozoan
filose amoeba closely related to the naked amoeboflagellate *Metopion fluens* (Bass et al.
2005). While the nature of the chlorarachniophyte endosymbiont remains a contentious
issue, the nature of the chlorarachniophyte host has been well established by molecular
phylogenetics.

1.6 The Cabozoans

A larger group including cercozoa and excavate protists has also been proposed
by T. Cavalier-Smith (Cavalier-Smith 2003), based on the premise that euglenids and
chlorarachniophytes share a common photosynthetic ancestor. The group containing these two lineages has been given the name Cabozoa, in recognition of the presence of chlorophyll a and b in the plastids of both chlorarachniophytes and euglenids. Both photosynthetic euglenids and chlorarachniophytes branch within larger groups of non-photosynthetic eukaryotes, the excavates and rhizaria, respectively.

The cabozoa hypothesis requires that excavates and rhizarians share a close common photosynthetic ancestor and that many independent plastid losses have occurred in both groups. Cavalier-Smith (2003) predicts that 3 losses of chloroplast within the Rhizaria and 6 within the Excavata would be required to assume a common evolutionary origin of chlorarachniophyte and euglenid plastids (Cavalier-Smith 2003). In support of the cabozoa hypothesis, Cavalier-Smith (2003) argues that “a tremendous economy in the mechanisms for importing nuclear-coded chloroplast proteins” (Cavalier-Smith 2003) would be provided by a single origin of a secondary green chloroplast. Rather than re-inventing chloroplast peptide-import via the Golgi, a simpler alternative would be to assume a common ancestor of both lineages with a common import system (and associated plastid-targeted peptides) already in place. Cavalier-Smith (2003) also notes other incidental similarities between chlorarachniophytes and euglenids, such as the shared use of cytosolic paramylon as a storage product (see above). Cavalier-Smith (2003) dismisses phylogenetic evidence (Deane et al. 2000) that fails to group chlorarachniophytes and euglenids together, arguing that plastid-targeted genes from euglenids and chlorarachniophytes both evolve too rapidly to place them accurately.

The primary objection to the Cabozoa hypothesis is the number of independent plastid losses required to explain the distribution of green plastids in rhizarians and euglenozoa, an issue acknowledged by Cavalier-Smith (Cavalier-Smith 2003), but counter-balanced by the improbability of re-inventing plastid-targeting twice in euglenids and chlorarachniophytes. Unlike the chromalveolates, no molecular evidence exists to support this group that is not subject to an alternate explanation, such as poor sampling or paralogy issues. Cavalier-Smith notes that the resolution of this debate will come only with a detailed understanding of the targeting machinery of both groups. Another resolution for this debate might center on similarities in chloroplast genome sequence in chlorarachniophytes and euglenids. Although chloroplast genome comparison cannot
prove the cabozoa hypothesis, as the chloroplasts of both lineages may have evolved from similar green algae, it can be used to argue against the cabozoa hypothesis if there are dissimilarities at the level of chloroplast genomes.

1.7 The Chromalveolates

Cavalier-Smith also proposed a relationships between red plastid containing organisms, cryptophytes, stramenopiles, haptophytes and alveolates. The chromalveolate hypothesis (Cavalier-Smith 1999) proposes that a single secondary endosymbiotic event involving a red algal cell gave rise to the plastids of all of the above groups; hence these groups are related at both the host and symbiont level. The chromalveolate hypothesis is complicated by the occurrence of non-photosynthetic lineages at the base of the cryptophytes, heterokonts, dinoflagellates and apicomplexans, and the absence of a plastid in ciliates.

Molecular phylogenies produce conflicting support for the chromalveolates. Concatenated analysis of nuclear encoded genes (Harper, Waanders, and Keeling 2005) recovers strong support for a clade of heterokonts and alveolates, and a clade of cryptophytes and haptophytes, but fails to group the two clades together. Molecular evidence for the chromalveolate theory comes chiefly from plastid gene replacement events. Photosynthetic heterokonts, haptophytes and dinoflagellates, as well as Toxoplasma gondii, have a plastid-targeted GAPDH that is clearly related to a cytosolic form of this enzyme in apicomplexans, dinoflagellates, ciliates and heterokonts (Fast et al. 2001). Fast et al., argue that the nuclear-encoded cytosolic GAPDH gene was duplicated early in chromalveolate evolution, the product of this duplicated gene became targeted to the plastid and replaced what would have been an ancestral red algal plastid-targeted gene (Fast et al. 2001). An expanded data-set of chromists including haptophytes recovered a similar relationship with all representative of the chromalveolates included (Harper and Keeling 2003). A second such event has been described by Patron et al. (Patron, Rogers, and Keeling 2004), who note that heterokonts, haptophytes, cryptophytes and dinoflagellates have a class of fructose-bisphosphat aldolase unrelated to the plastid-targeted form of this enzyme in red algae, suggesting again that a replacement of the ancestral red algal plastid-targeted enzyme has occurred
early in chromalveolate evolution. Although the balance of molecular evidence supports the chromalveolate theory, the status of this group remains contentious.

1.8 Structure and objectives of thesis

Chlorarachniophytes provide an excellent system in which to study the effects of secondary endosymbiosis on eukaryotic cells. The objective of this thesis is to examine the effects of secondary endosymbiosis on the chlorarachniophyte cell from several different perspectives. Although these chapters discuss disparate topics, the underlying theme behind all of these different topics is an emphasis on gene transfer. Previous studies have demonstrated that many genes of nucleomorph descent have been transferred to the host nucleus and are targeted back to the plastid (Deane et al. 2000; Archibald et al. 2003; Gilson et al. 2006). Although the nucleomorph genome encodes proteins functioning in the stroma (Gilson et al. 2006), this is a sparse representation of known nuclear encoded genes in green algae and plants. The bulk of these ancestrally nucleomorph encoded genes have been transferred to the host nucleus, or have perhaps been replaced by genes of host origin that have acquired presequences for entry into the endomembrane system and subsequently the plastid. Chapter two discusses two examples of possible endosymbiotic gene transfers from the endosymbiont to the host. One of these involves the plastid-targeted class I fructose bisphosphate aldolase of *B. natans* which shares a weakly supported relationship with a divergent green algal class I FBA. The plastid-targeted FBPase and SBPase of *B. natans* both branch with other plastid-targeted forms of these enzymes in other eukaryotes, though their relationships are equivocal, thus it cannot be conclusively said that they originated via endosymbiotic gene transfer rather than lateral gene transfer from a different photosynthetic eukaryote. In addition to describing these three putative endosymbiotic gene transfer events involving chlorarachniophytes, chapter two also touches on a variety of other gene transfer and replacement events involving the non-homologous but functionally equivalent class I and class II fructose bisphosphate aldolase. Class II aldolase is likely the ancestral plastid form of this enzyme, it is present in a diversity of cyanobacteria, as well as the glaucocystophytes *Cyanophora paradoxa* which appears to have a class II aldolase directly descended from these cyanobacteria (Nickol et al. 2000). Green algae, plants and
red algae have replaced their class II aldolase with a plastid-targeted class I aldolase which may be derived from a cytosolic eukaryotic-type class I aldolase (Gross et al. 1999). Chromists have a plastid-targeted class II aldolase unrelated to that of Cyanophora paradoxa, suggesting a further replacement of the ancestral red algal class I aldolase in these organisms. Implications of this observation for the chromalveolate hypothesis are discussed by Patron et al. (2004) who include a wider variety of chromistan class II FBAs in their analysis. Gene transfer is also the subject of chapter three, this chapter deals with lateral transfer of genes between unrelated eukaryotes and prokaryotes. Many lateral gene transfer events between chlorarachniophytes and prokaryotes and other phototrophic eukaryotes have been previously described (Archibald et al. 2003). As mixotrophic cells, chlorarachniophytes rely in part on a heterotrophic lifestyle. Previous studies have indicated that chlorarachniophytes are capable of consuming a vast array of different cell types, both eukaryotic and prokaryotic (Hibberd and Norris 1984; Moestrup and Sengco 2001). This phagotrophic lifestyle has created an abundance of opportunities for new genes to be acquired through lateral gene transfer. This has been shown to have occurred multiple times in the history of B. natans from a variety of different sources, both prokaryotic and eukaryotic (Archibald et al. 2003). Chapter three discusses Calvin cycle and pentose phosphate enzymes that have arisen through eubacterial gene transfer events. Two of these have previously been characterised by Archibald et al. (2003). We show that with increased sampling of additional eukaryotic taxa a gamma-proteobacterial like enzyme in B. natans branches with chromistan sequences suggesting that transfer of this gene between eukaryotes has occurred following a prokaryote to eukaryote transfer. A relationship between a laterally transferred GAPDH present in diplonemids and chromistan algae is also suggestive of gene transfers between eukaryotes following a prokaryote to eukaryote transfer event. A final example presented in this chapter is the transfer of a chlamydiales-like transketolase gene among several groups of unrelated eukaryotes. Chapter four is adapted from a published manuscript (Rogers et al. 2004) and discusses the nature of the plastid-targeting peptides in B. natans. Understanding targeting is an essential component of explaining endosymbiotic gene transfer as genes transferred to the nucleus of their host are commonly directed back to the plastid through the acquisition of N-terminal targeting presequences. This chapter describes the targeting
pre-sequences of plastid-targeted proteins in chlorarachniophytes and how they compare to these sequences in other eukaryotes where they have been studied. In addition, a heterologous targeting experiment using the apicoplast of *Toxoplasma gondii* demonstrates that signal peptides of chlorarachniophyte plastid-targeted peptides are sufficient for import into the endomembrane system of this apicomplexan but that targeting of these proteins to the apicoplast does not occur. The results of this experiment suggest that a difference exists in plastid-targeting between apicomplexans and chlorarachniophytes either at the level of the transit peptide sequence or transit peptide recognition by the inner membranes of the plastids of these organisms. Chapter five describes the completed chloroplast genome of *B. natans*. The chloroplast genome is the ultimate source of all endosymbiotic gene transfers and many inferences have been made about plastid origins in different groups of photosynthetic eukaryotes based on shared losses of chloroplast genes. The plastid genome of *B. natans* is the smallest sequenced plastid genome from a photosynthetic organism. Despite its reduced size, this plastid genome encodes a complement of protein coding gene typical of photosynthetic green algal genomes. Comparison of coding DNA content between *B. natans* and other sequenced genomes indicates that although the chloroplast genome of *B. natans* contains less coding DNA than any other plastid genome from a phototrophic eukaryote, the difference in size between it and chloroplast genomes of green algae, plants and euglenids is primarily due to the absence of introns and non-coding sequence. Phylogenetic analyses of protein coding genes encoded in the plastid genome of *B. natans* suggest that the endosymbiont of *B. natans* is related to UTC green algae, precluding a prasinophycean, streptophyte or basal chlorophyte origin of the chlorarachniophyte symbiont and presenting strong evidence against Cavalier-Smiths Cabozoa hypothesis.
1.9 Bibliography


Ishida, K., T. Nakayama, and Y. Hara. 1996. Taxonomic studies on the Chlorarachniophyta. II. Generic delimitation of the chlorarachniophytes and


Chapter 2 - Lateral Transfer and Recompartementalisation of Calvin cycle enzymes of plants and algae*

2.1 Introduction

Plastids, the photosynthetic organelles of plants and algae, arose through the endosymbiotic uptake of a cyanobacterium. Primary plastids, found in glaucophytes, red algae, green algae and plants are the direct descendants of the original symbiosis between a cyanobacterium and a non-photosynthetic eukaryote. All other plastids have arisen through an endosymbiotic partnership between one of these groups of algae and a non-photosynthetic eukaryote. These plastids are referred to as secondary plastids. Ultimately, all plastids trace back to the cyanobacterial endosymbiont, so plastid biochemistry is typically performed by cyanobacterial enzymes, even though the genes for most of these enzymes have been relocated to the nuclear genome of the host alga. Almost all Calvin cycle reactions (or reverse reactions) are also used in glycolysis, gluconeogenesis, or the pentose phosphate pathway, and many of these are carried out by homologous enzymes in the host cytosol. In these cases, the ancestor of plants and algae would have contained two genes with distinct evolutionary histories, one cyanobacterial gene for a plastid-targeted protein and a second eukaryotic gene for the cytosolic enzyme. In several cases, this simple prediction is apparently not met, as gene duplications, losses, and transfers all seem to have taken place, blurring any hard line between host and endosymbiont at the molecular level. Two Calvin cycle enzyme families that do not conform to this prediction

* This chapter has been published as a manuscript: Rogers, M. B., and P. J. Keeling. 2003. Lateral gene transfer and re-compartmentalisation of Calvin cycle enzymes in plants and algae. J. Mol. Evol. 58:367-375.
are fructose bisphosphate aldolase (FBA) and sedoheptulose bisphosphatase/fructose bisphosphatase (SBPase/FBase).

FBA is divided into two classes of phylogenetically and structurally unrelated enzymes characterised by several divergent properties (Marsh and Lebherz 1992). Class I FBAs are homotetramers that form a Schiff base with their substrate, and can be inhibited with borohydride reagents (Rutter 1964; Lebherz and Rutter 1969). Class II FBAs, on the other hand, are found as homodimers and require divalent cations as cofactors and are inhibited by EDTA (Rutter 1964; Zgiby et al. 2000). Both classes of FBA catalyze the cleavage of fructose-1,6-bisphosphate to glyceraldehyde-3-phosphate and dihydroxyacetone phosphate in glycolysis or the reverse aldol condensation in the Calvin cycle. In addition to the aldol condensation of triose sugars, the plastid FBA I of Spinacia can condense DHAP and erythrose-4-phosphate to form sedoheptulose-1,7-bisphosphate (Brooks K 1966). The taxonomic distribution of both classes is complex and probably not fully realized. Class I FBA is found primarily in eukaryotes, where it is widespread (Rutter 1964; Lebherz, Leadbetter, and Bradshaw 1984; Marchand et al. 1988; Jacobshagen and Schnarrenberger 1990; Pelzer-Reith, Penger, and Schnarrenberger 1993; Pelzer-Reith, Wiegand, and Schnarrenberger 1994; Schnarrenberger et al. 1994; Plaumann et al. 1997; Gross et al. 1999). A small group of proteobacteria also possess class I FBA, the role of which is unclear, and archaeabacteria and various Gram-positive bacteria possess distinct FBAs that are distantly related to the FBA I of eukaryotes and proteobacteria (Siebers et al. 2001). Red algae, green algae and plants possess two distinct class I FBAs, one glycolytic and gluconeogenic enzyme that functions in the cytosol and a second, plastid-targeted Calvin cycle enzyme, which is thought to have arisen by gene duplication (Kruger and Schnarrenberger 1983; Gross et al. 1999). Class II FBA is chiefly found in eubacteria, where there is extensive paralogy, and is divided into two distinct subgroups, type A and type B. The two types are distantly related and are distinguished by several large insertions and deletions (Plaumann et al. 1997; Henze et al. 1998; Nickol et al. 2000). Type A FBA II has been characterised from a variety of eubacteria, as well as the cytosol of ascomycete fungi and Euglena (Rutter 1964; Marsh and Lebherz 1992; Pelzer-Reith, Wiegand, and Schnarrenberger 1994; Plaumann et al. 1997). Type B FBA II is also present in diverse eubacteria, several “amitochondriate”

FBPase and SBPase are related enzymes that catalyze similar reactions in the Calvin cycle: SBPase catalyzes the substrate level dephosphorylation of sedoheptulose-1,7-bisphosphate to sedoheptulose-7-phosphate (Cadet and Meunier 1988), while FBPase catalyzes the dephosphorylation of fructose-1,6-bisphosphate to fructose-6-phosphate (Zimmermann, Kelly, and Latzko 1976). FBPase is widely distributed, and the plastid FBPase of red algae, green algae and plants are not cyanobacterial but instead are related to cytosolic forms, suggesting they originated through gene duplication (Martin et al. 1996). In contrast, SBPase is found only in plastids of green algae and plants and the kinetoplastid, *Trypanosoma*. Like FBA, the presence of an SBPase in *Trypanosoma* is a key element in the argument for an ancestral plastid in kinetoplastid parasites (Hannaert et al. 2003).

We have investigated the evolution of plastid and cytosolic FBA, FBPase and SBPase genes by reconstructing the phylogenies of all three enzymes including plastid and cytosolic sequences from algae with secondary plastids derived from both green and red algae. The evolution of plant and algal FBA, FBPase and SBPase is significantly more complex than previously considered, and proteins targeted to the plastids of plants and algae have a variety of evolutionary histories.

2.2 Methods

2.2.1 Identification, sequencing, and primary analysis of FBA and FBPase/SBPase genes

Messenger RNAs encoding class I and class II FBA as well as SBPase and FBPase were identified from an ongoing *Bigelowiella natans* EST sequencing project (www.botany.ubc.ca/keeling/ChlorEST/ChlorEST.html) based on sequence similarity to known eukaryotic and eubacterial FBA genes. All EST clones matching any known FBA, FBPase or SBPase were isolated and fully sequenced as described (Archibald et al. 2003). Inferred amino acid sequences from all algal genes were analysed for probable cellular location by identifying and characterising any putative N-terminal targeting
information. Secondary plastids such as those in *Bigelowiella*, diatoms, apicomplexa and *Euglena* use a bipartite leader consisting of a signal peptide followed by a transit peptide, both of which have several predictable chemical characteristics (McFadden 1999). All genes from these organisms were examined for the presence of a peptide leader, which were analysed using SignalP and IPsort to identify putative signal and transit peptides, respectively (Nielsen et al. 1997; Bannai et al. 2002).

### 2.2.2 Phylogenetic Analysis

Available FBA, FBPase and SBPase were aligned using ClustalX X (Jeanmougin et al. 1998). ESTs encoding an SBPase like protein from *Magnaporthe grisea* were retrieved from the Whitehead Institute *Magnaporthe* EST databases (www-genome.wi.mit.edu and www.fungalgenomics.ncsu.edu). Trees were inferred from amino acids using distance and protein maximum likelihood methods. Distances were calculated by TREE-PUZZLE 5.0 (Strimmer and von Haeseler 1996) using the WAG substitution matrix (Goldman and Whelan 2000), and correcting for rates-across-sites variation according to a discrete gamma distribution with eight categories of variable sites and an invariable sites category. The shape parameter alpha and the proportion of invariable sites were estimated from the data. Trees were inferred from gamma-corrected distances by weighted neighbor joining using WEIGHBOR 1.0.1a (Bruno, Socci, and Halpern 2000). Bootstrapped distances were calculated by PUZZLEBOOT (A. Roger and M. Holder, http://www.tree-puzzle.de) under described conditions with the alpha and proportion of invariable sites parameters estimated from the original data. Protein maximum likelihood trees were also inferred for class I FBA, class II type A FBA, and class II type B FBA, FBPase and SBPase using ProML 3.6a (Felsenstein 1993) with global rearrangements and 10 randomized sequence additions. Rates-across-sites heterogeneity was modeled using the R-option with seven rate categories and frequencies estimated by TREE-PUZZLE (six variable and one invariable categories). Bootstrap resampling was performed using protein maximum likelihood in the same way, except that four variable rate categories were used. As both type A and type B class II FBA contain many insertions and deletions with respect to one another, otherwise informative characters are unalignable between the two subclasses. In order to optimize the character
sets of both types of class II FBA, and to more precisely resolve the phylogenies of type A and type B class II FBA individually, class II FBA was analysed in three steps. Initially, the phylogeny of both type A and type B together (using 270 alignable characters) was inferred and bootstrapped using distance methods. Then, the phylogenies of type A and type B were inferred independently using distance and maximum likelihood methods (using 324 and 267 characters, respectively). The phylogeny shown is based on the protein maximum likelihood trees of the two individual types, joined at the nodes found in the phylogeny of the entire class.

2.3 Results and Discussion

2.3.1 Class I FBA

The phylogeny of class I FBA (Figure 2.1) shows a number of well-supported groups, including apicomplexans, animals, proteobacteria, and both plastid-targeted and cytosolic forms from plants. The fact that both plastid and cytosolic FBAs in plants are class I enzymes, and that the plastid enzyme shows no resemblance to known cyanobacterial enzymes has been interpreted as evidence for an endosymbiotic gene replacement event in red algae, green algae and plants. It is thought that the gene for the cytosolic enzyme duplicated and the protein product of one paralogue was targeted to the plastid, leading to the deletion of the original cyanobacterial enzyme (Gross et al. 1999). Plastid-targeted proteins from three green algae are also strongly related to the plant plastid clade, while the plastid-targeted protein from the red alga Galdieria is also weakly related to the plant clade. Oddly, the Euglena plastid-targeted FBA shows no phylogenetic affinity to other plastid-targeted enzymes, green algal enzymes, or the glycosome-targeted FBA of its kinetoplastid relative, Trypanosoma. The origin of the Euglena plastid enzyme is accordingly uncertain (Plaumann et al. 1997), but seems unlikely to be derived from green algal plastid genes, as would be expected. The cytosolic and plastid-targeted class I FBAs of Bigelowiella branch weakly with their respective cytosolic and plastid-targeted counterparts in algae and land plants, as would be expected.
Figure 2.1 Protein maximum likelihood tree of class I FBA

Numbers at nodes correspond to bootstrap values greater than 50% for major nodes obtained from weighted neighbor-joining (top), and protein maximum likelihood (bottom). Plastid-targeted genes are in black boxes. Major groups are indicated by brackets and labelled.
Interestingly, the *Bigelowiella* plastid-targeted FBA branches with poor support (45% ProML bootstrap support) with a divergent FBA from *Chlamydomonas* that encodes a leader predicted to be a plastid transit peptide. It appears that *Bigelowiella* has retained this divergent type of green algal FBA in its plastid.

A recent analysis of the class I FBA of *Trypanosoma* has been used to suggest that kinetoplastids have acquired their glycosomal FBA enzyme through an ancient endosymbiosis with an alga (Hannaert et al. 2003). However, we do not find this relationship when a broad sampling of eukaryotes and proteobacterial FBAs are included in the analysis (several key taxa such as *Euglena*, *Dictyostelium*, *Cryptosporidium*, and proteobacteria were not included in that analysis). Rather, the position of *Trypanosoma* in the tree is equivocal, but it never branches with plastid-targeted FBAs. Consequently, there is no evidence that it is derived from a plastid-targeted FBA. Likewise, the FBA of alveolates shares no close affinity with that of algae and plants, as has also been suggested (Hannaert et al. 2003).

**2.3.2 Class II FBA**

The composite phylogeny of type A and B subgroups of class II FBA is shown in Figure 2.2. Both types are predominantly composed of eubacterial sequences, where paralogy is apparent in several cases, and the overall relationships are not well supported. Eukaryotic and plastid-targeted class II FBA genes appear in various places in the tree. The cytosolic class II type A FBA of *Bigelowiella* falls within a clade of eukaryotic genes, and shares a robustly supported branch with the cytosolic FBA of *Euglena*. This relationship is intriguing as both organisms contain secondary plastids of green algal origin, but share few other features that would support a close common line of nucleocytoplasmic descent. Nevertheless, a relationship between euglenids and chlorarachniophytes (and many other eukaryotes) has been proposed based on the proposition that their plastids originated through a single common secondary endosymbiotic event (Cavalier-Smith 1998; Cavalier-Smith 2000). However, the kinetoplastid *Trypanosoma* has a class I FBA, but no class II FBA is present in the unfinished genomes of *Trypanosoma brucei* or *Leishmania major*. If the *Euglena* and
Bigelowiella class II FBAs were both inherited from a common ancestor, then the trypanosomatids should also have inherited this enzyme. Alternatively, both may have acquired the enzyme (in parallel or in common) from their endosymbiont, but once again, no class II FBA sequences have been identified from chlorophytes including the complete genome of Chlamydomonas. The very strong relationship between the cytosolic FBAs of Euglena and Bigelowiella lacks an obvious explanation, but it is nonetheless an intriguing link between these two lineages. The position of the Euglena and Bigelowiella enzymes within the larger picture of class II type A FBAs is also of interest. Formerly only Euglena and fungi were known to have genes of this class, but now this strongly supported clade also includes Bigelowiella and a cytosolic homologue (based on the absence of an N-terminal extension and a methionine start codon at approximately the same position as the confirmed cytosolic FBA of fungi) assembled from ESTs from the oomycete heterokont Phytophthora sojae.

A second, strongly supported, grouping of FBAs from the diatoms Phaeodactylum tricornutum and Odontella sinensis consistently branch at the base of this clade with modest support. Like chlorarachniophytes and euglenids, these algae possess secondary plastids (in this instance of red algal origin), and these proteins have bipartite leaders with signal and transit peptide moieties (one of the Phaeodactylum FBAs is truncated within a partial leader sequence), indicating they are plastid-targeted. The two Phaeodactylum paralogues are not specifically related, indicating a gene duplication sometime in the ancestry of diatoms. Since red algae have both cytosolic and plastid class I FBA, the plastid targeted class II FBA of diatoms could not have been acquired from their red algal endosymbiont. Given the existence of cytosolic class II FBA in a variety of other eukaryotes including an oomycete, it is possible that the plastid targeted class II FBA of diatoms originated from the targeting of a host cytosolic enzyme to the plastid.
Figure 2.2 Composite protein maximum likelihood tree of type A and type B class II FBA

Type A and B trees were individually constructed are rooted (indicated by a dotted line) as described in the Methods section. Numbers at nodes of subtrees correspond to bootstrap values greater than 50% for major nodes obtained from weighted neighbor-joining (top) and protein maximum likelihood (bottom). Bootstrap values in italics correspond to weighted neighbor-joining values for the position of the root in global analyses. Plastid-targeted and other eukaryotic enzymes are shaded and boxed respectively, as in Figure 2.1.
Several eukaryotes possess class II type B FBAs, but unlike type A, these are scattered about the type B subtree. These are the plastid-targeted FBA of the glaucophyte *Cyanophora paradoxa* (discussed below), the amitochondriates, *Giardia*, *Spironucleus*, *Trichomonas*, and *Mastigamoeba*, and poorly characterised land plant genes that were identified in genome sequencing data or assembled from ESTs. Neither of the two type B sequences from *Arabidopsis* or *Glycine* are predicted to encode an N-terminal transit peptide, indicating that these enzymes are localised in the cytosol. In some analyses they show a weak affinity for the gamma proteobacterial genes for tagatose bisphosphate aldolase and an uncharacterised gene from the gamma proteobacteria *Yersinia pestis*. This probably reflects a bacterial origin and perhaps a non-fructose substrate for these enzymes. *Cyanophora paradoxa* has a class II type B FBA that is unrelated to any of the other eukaryotic homologues, but occupies a position of special interest. This gene encodes a plastid-targeted FBA, which branches with cyanobacterial FBAs with very strong support (Nickol et al. 2000). Normally, a plastid-targeted enzyme of cyanobacterial affinity would not be particularly interesting, but in this case the plastid enzymes of red and green algae are class I enzymes likely derived from endosymbiotic gene replacement (Gross et al. 1999). This key difference is potentially important evidence for the early divergence of glaucophytes, since they are the only algae that possess the ancestral plastid-targeted FBA. A deep position of glaucophytes has been debated extensively (e.g., (Cavalier-Smith 1982; Bhattacharya et al. 1995; Helmchen, Bhattacharya, and Melkonian 1995; Martin et al. 2002)), but there is not yet conclusive evidence either way. The FBA gene replacement may be quite helpful in supporting the glaucophytes as the earliest lineage of primary algae because it is possible to infer if the replacement took place once or twice independently based on the phylogeny of class I FBA. If an ancestral class II FBA has been replaced by a class I FBA once in a common ancestor of red algae and green algae, then class I FBA phylogeny should show the plastid and cytosolic genes of red and green algae forming distinct clades. Conversely, if red and green algae independently replaced their plastid class II enzymes with class I enzymes, then the plastid and cytosolic enzymes from each algal group should be most closely related to each other. Unfortunately the overall level of support seen in class I
FBA phylogeny is low, but it does consistently support the former alternative in all analyses (figure 2.1). It therefore appears most likely that the gene plastid FBA replacement took place once in the common ancestor of green and red algae, and by extension that glaucophytes diverged prior to this event.

### 2.3.3 FBPase/SBPase

The evolution of plastid-targeted FBPase in plants and algae is similar to that of FBA I: the plastid-targeted FBPase likely arose through duplication of an ancestral cytosolic FBPase (Martin et al. 1996). The plastid-targeted FBPase of *Bigelowiella* branches with other plastid-targeted FBPase genes and it also shares a 12 amino acid insertion with them. This insertion is absent in cytosolic FBPase, and has been shown to contain amino acids involved in thioredoxin binding (Hermoso et al. 1999). The second *Bigelowiella* FBPase is equivocal in its position in the tree, it branches with low support at the base of the clade consisting of eukaryotic FBPase in a manner analogous to the cytosolic FBA I of *Bigelowiella*. In contrast, the cytosolic FBPase of the apicomplexan *Toxoplasma gondii* branches within a larger clade of eubacteria, consisting of a cyanobacterium, *Chlorobium* and several gamma-proteobacteria. This relationship is well-supported and consistent with a eubacterial origin of this enzyme, although it is unclear which group of eubacteria may have served as a donor. Additional ESTs from two other coccidians, *Eimeria tenella* and *Neospora caninum* (GENBANK accession #s: BI895768; BF248534) are highly similar to the *Toxoplasma* sequence, but no indications of an FBPase are present in genomic sequences from *Plasmodium*, *Cryptosporidium*, or *Theileria*. Altogether, this suggests that a eubacterial FBPase has most likely been transferred to a common ancestor of coccidians following their divergence from other apicomplexans. This transfer is of functional interest, as it suggests that an important differential core carbon metabolic capability exists between different apicomplexans that was enhanced by lateral transfer. The evolutionary history of other gluconeogenic enzymes for which little information from coccidians is available (e.g., glucose-6-phosphatase) would be interesting to examine. In plants and algae that have been examined, SBPase is exclusively found in the plastid and has no cytosolic role. Accordingly, the characterisation of a cytosolic SBPase in the non-photosynthetic
kinetoplastid *Trypanosoma brucei* led to the suggestion that this enzyme (like FBA) was derived from a cryptic plastid endosymbiont (Hannaert et al. 2003). However, Figure 2.3 shows that the fungi *Magnaporthe grisea* and *Neurospora crassa* also possess an apparently cytosolic SBPase, and this fungal gene forms a clade with the kinetoplastid SBPase that receives good bootstrap support from both maximum likelihood and distance methods (we also analysed these data using Fitch-Margoliash where the position of *Trypanosoma* was equivocal: not shown). On the whole, it appears that SBPase may have been a late addition to plastid metabolism, having persisted in the cytosol of non-photosynthetic eukaryotes prior to the initial acquisition of the chloroplast. Indeed, cyanobacteria lack SBPase and use a dual specificity FBPase (Yoo and Bowien 1995; Tamoi et al. 1996), so the plastid SBPase most likely came from some source other than cyanobacteria. Accordingly, there is no reason to believe that SBPase in non-photosynthetic protists are indicative of a plastid. Still, the function of SBPase in *Trypanosoma* and fungi is puzzling. Hannaert et al. (2003) suggest that the SBPase of *Trypanosoma* may function in a modified pentose phosphate pathway. This would require the action of an FBA I specific for erythrose to generate sedoheptulose-1,7-bisphosphate, as is thus far only known to occur in the plastids of plants (Brooks K 1966) and alphaproteobacteria. In light of these findings, experimental evidence for the substrate specificity of fungal and kinetoplastid FBPase would prove exceedingly interesting.
Figure 2.3 Protein maximum likelihood phylogeny of FBPase and SBPase.
Numbers at nodes, shading and brackets are as described in Figure 2.1.
2.4 Concluding remarks: the evolution of FBA, FBPase and SBPase.

The evolutionary history of enzymes in eukaryotic core carbon metabolism has frequently been quite colourful, and the redundancy of enzymes resulting from the endosymbiotic origin of the plastid has added to this in several interesting ways. In the cases of the three enzymes examined here, we find instances of lateral gene transfer, re-targeting of enzymes between cellular compartments, and instances where one or both of these processes support important relationships in the tree of eukaryotes. Two strong cases of lateral transfer are found in the plant class II type B FBA and the coccidian FBPase, both of which have interesting functional implications for the metabolism of the recipient. The diverse group of eukaryotic class II type B FBA genes might also have arisen by one or more lateral transfer events, but this is not so clear. The re-targeting of an enzyme to a new cellular compartment has been observed previously in both FBA (plants and primary plastid containing algae) and FBPase (plants), and we now also show that it most likely accounts for plastid SBPase and perhaps also the heterokont plastid class II type B FBA. While characterisation of such events is becoming more common (Brinkmann and Martin 1996; Fast et al. 2001), they may still provide important evidence for major events or lineages in the tree of eukaryotes. In this case, the retention of an ancestral class II type A FBA in the plastid of *Cyanophora* is a strong indication that glaucophytes diverged prior to green and red algae.
2.5 Bibliography


Jacobshagen, S., and C. Schnarrenberger. 1990. 2 class-I aldolases in Klebsormidium flaccidum (Charophyceae) - an evolutionary link from chlorophytes to higher-Plants. J. Phycol. 26:312-317.


Chapter 3 - A complex and punctate distribution of three genes derived by lateral gene transfer in eukaryotic genomes.*

3.1 Introduction

Lateral gene transfer is the movement of genes between distantly related organisms, and has become a major focus in the study of genome evolution (Doolittle 1999; Eisen 2000; Ochman, Lawrence, and Groisman 2000; Koonin, Makarova, and Aravind 2001; Gogarten, Doolittle, and Lawrence 2002; Beiko, Harlow, and Ragan 2005; Coleman et al. 2006; Ragan, Harlow, and Beiko 2006). The importance of gene transfers between prokaryotic genomes is now generally recognized due to the extensive availability of prokaryotic genomes for comparison, although there is still controversy about how common it is, or what long term effects it has (Lawrence and Hendrickson 2003; Kurland 2005; Hao and Golding 2006). For eukaryotes there are far fewer genome sequences to compare and emerging evidence that it may not be prevalent in many of the lineages where the most data are available, such as vertebrates (Salzberg et al. 2001; Stanhope et al. 2001). Nevertheless, an abundance of convincing examples of prokaryote-to-eukaryote gene transfer events have been described (Boucher and Doolittle 2000; Nixon et al. 2002; Andersson et al. 2003; Archibald et al. 2003; Waller, Slamovits, and Keeling 2006), and transfers between eukaryotes are also known (Keeling and Palmer 2001; Archibald et al. 2003; Bergthorsson et al. 2003; Takishita, Ishida, and Maruyama 2003; Andersson 2005). While these make it clear that lateral transfer has affected eukaryotic nuclear genomes, the extent of its impact remains unknown. In particular, most known cases involve genes moving from a prokaryote to eukaryotes, whereas comparatively little is known about transfers between eukaryotes.

Various methods have been used to infer transfers between prokaryotic genomes (Ragan 2001), but by far the most common method of detecting events involving

* This chapter is a manuscript in preparation: Rogers MB, Watkins RF, Harper JT, Durnford DG, Gray MW and Keeling PJ. A complex and punctate distribution of three genes derived by lateral gene transfer in eukaryotic genomes.
eukaryotes is to observe incongruence between phylogenetic trees based on a gene and the tree we believe to reflect the evolution of the organism in which the gene is encoded. Lateral gene transfer events are thus commonly invoked to explain trees that depart from expectations. Nevertheless, other explanations can account for these incongruent topologies, including reconstructions that do not accurately reflect the history of the gene due to problems like the failure to account for rate across site variation, or covariance and biased amino acid composition across the tree (Roger and Hug 2006). Similarly, the history of the gene may be complex in ways that erroneously suggest lateral transfer even when the phylogeny is accurately reconstructed. For example, gene duplication events and differential extinction of the paralogues across different lineages can lead to a tree that appears to reflect lateral transfer but really describes the history of a duplicated gene.

Another emerging problem for the interpretation of lateral gene transfer are genes with a punctate distribution. For eukaryotes, this has been used to refer to cases where two or more distantly related eukaryotic lineages possess closely related genes that are either not found in other eukaryotes, or are clearly different from other eukaryotic homologues (Keeling and Inagaki 2004). This is in contrast to the term ‘patchy’, which has been used to refer to genes with a limited distribution in both eukaryotes and prokaryotes (Andersson et al. 2006). Genes with a punctate distribution are significant because they have been interpreted in several different ways, each with its own important implications. On one hand, such distributions have been interpreted to represent a single transfer to the common ancestor of the two or more disparate lineages. The reconstruction of several large-scale relationships among eukaryotes (the so-called supergroups) has been partially based on the documentation of shared, rare characteristics like gene fusions or indels in two or more lineages (Baldauf and Palmer 1993; Archibald et al. 2002; Stechmann and Cavalier-Smith 2002). Shared lateral gene transfers have also been used in this way, for example the shared possession of a nanoarchaeal prolyl-tRNA synthetase in trichomonad and diplomonad flagellates (Andersson, Sarchfield, and Roger 2005) or a haloarchaeal tyrosyl-tRNA synthetase in opisthokonts (Huang, Xu, and Gogarten 2005). On the other hand, more complex distributions have been interpreted to represent multiple transfers or eukaryote-to-eukaryote transfers: such seems to be the case for the recently described EF1-alpha (EFL) like GTPases of eukaryotes (Keeling and Inagaki...
2004). These are also significant because detecting transfers between eukaryotes is made difficult by poor sampling and poor resolution of many phylogenies, both of which impede the distinction between horizontal and vertical descent. For this reason, some eukaryote-to-eukaryote transfers have been argued based on the gene bearing some special characteristic that makes it stand out, such as an insertion, an accelerated rate of substitution, or even an origin from another lateral transfer event (Keeling and Palmer 2001; Keeling and Inagaki 2004; Andersson, Sarchfield, and Roger 2005).

For any of these genes the possibility of ancient paralogy must also be considered, in which case the origin of the gene may still ultimately be due to a lateral transfer event, but its distribution is due to other factors. Deciding between these interpretations is based on balancing a variety of observations, including how widespread the gene in question is, how closely related we believe the organisms that possess the gene to be, whether other close relatives possess or lack the gene, and how closely related the gene is to some identifiable source. In many cases we cannot answer these questions because the data are lacking from a sufficient diversity of eukaryotes, so it is impossible to conclude what might be the cause of known cases. Moreover, it is also impossible to say whether the processes leading to punctate clades are common or rare: in many cases bacterium-to-eukaryote transfers have been inferred from data with simple distributions, but it is possible that some of these only appear simple because of the lack of sampling.

Here we use EST data to evaluate several cases of apparently simple lateral gene transfer of genes involved in core carbon metabolism, and find each in each a more complex, punctate distribution than appreciated from the initial observations. We characterized protist homologues of three genes, ribulose-5-phosphate-3-epimerase (RPE), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), and transketolase. In the case of ribulose-5-phosphate epimerase, a transfer event has been described between a gamma-proteobacterium related to \textit{Pseudomonas} and a chlorarachniophyte (Archibald et al. 2003). Similarly, many lateral transfer events involving eukaryotic GAPDH have been described (Markos, Miretsky, and Muller 1993; Wiemer et al. 1995; Figge et al. 1999; Qian and Keeling 2001; Archibald et al. 2003; Takishita, Ishida, and Maruyama 2003), including one isolated clade of diplonemid genes inferred to be transferred from a proteobacterium (Qian and Keeling 2001). In the case of transketolase, the
chlorarachniophyte cytosolic transketolase is not closely related to other eukaryotic homologues, but is very similar to homologues from Chlamydiales. For each of these genes, a single transfer from a bacterium to a eukaryote is evident from the strong relationship of the eukaryotic gene to a particular subgroup of bacteria, but here we show that the distribution within eukaryotes suggests a more complicated history. In all three cases we find homologues in other eukaryotes that are distantly related to the organism where the gene was first found. How the complex distribution of these genes arose is uncertain, but these examples indicate that such a punctate pattern of presence is more common than previously thought, and that other apparently simple cases of lateral transfer may be more complex than they appear.

3.2 Methods

3.2.1 Characterisation of new sequences

Clones corresponding to transketolase from *Euglena gracilis*, *Hartmanella vermiformis*, *Karlodinium micrum*, *Isochrysis galbana* and *Physarum polycephalum* were identified from TBestDB (http://amoebidia.bcm.umontreal.ca/pepdb/), and re-sequenced to produce assemblies encoding a full length coding sequence where possible. Full-length assemblies were assembled in this way from *P. polycephalum* and putative full-length assemblies were obtained for *H. vermiformis* and *E. gracilis*. 5’-truncated assemblies were obtained from *K. micrum* and *I. galbana*. The *B. natans* transketolase was assembled from three ESTs and the 5’ end was obtained through PCR amplification using a forward degenerate transketolase primers (CGCGACTACAGGGCCNYTNGGNCARGG) and a specific reverse primer based on the EST sequences (CTCTCCAACACCCGATAGAATCATGAGTC). A second degenerate PCR product encoding an upstream region of the *B. natans* transketolase was obtained using a second degenerate primer (CGCGACTACAGGGCCNYTNGGNCARGG) and a specific reverse primer based on the above PCR product (GCCTTGTACCGGGTGATGACATCCTCAG). A PCR product from *Isochrysis galbana* with similarity to bacterial GAPDHs was obtained through PCR using degenerate primers (CCAAGGTCGGNATHAAYGGNTTY and
CGAGTAGCCCCAYT CRTTRTCRTACCA). All PCR products were cloned using the TOPO TA vector (Invitrogen) and multiple copies sequenced on both strands. Homologues of all three genes were also identified in the completed genomes of *Thalassiosira pseudonana* (Armbrust et al. 2004) and *Phaeodactylum tricornutum* (http://genome.jgi-psf.org/Phatrl/Phatrl.home.html). Updated models of these genes were provided by BR Green.

### 3.2.2 Sequence analyses

To evaluate the probable cellular location of each protein, the N-terminal regions of all full-length inferred protein sequences were analysed using SignalP v. 3.0 for the presence of an N-terminal signal peptide (Bendtsen et al. 2004) and leader sequences were examined for characteristics expected of transit peptides for the group in question.

New sequences were aligned to homologues from public databases using CLUSTAL X, and manually edited using MacClade 4.07. Publicly available sequences used in alignments were downloaded from GenBank, ESTdb, or from complete eukaryotic genome databases. The full transketolase alignment included 473 characters, and a reduced alignment with 212 characters was also used to include those sequences with missing a large amount of 5′ sequence. The ribulose-5-phosphate epimerase alignment consisted of 183 characters and the GAPDH alignment consisted of 278 characters.

Phylogenetic analyses were carried out using maximum likelihood, distance and Bayesian methods. Maximum likelihood phylogenies were performed using PHYML 2.4.4. (Guindon and Gascuel 2003) with the WAG substitution matrix and site rate distribution modeled on a discrete gamma distribution with 4 rate categories and one category of invariable sites. The estimated alpha parameters were 1.154, 1.156, and 1.419 and the estimated proportion of invariable sites were 0.067, 0.050, and 0.098 for RPE, GAPDH, and transketolase, respectively. Bayesian analysis of trees was performed using Mr. Bayes 3.0b4 (Ronquist and Huelsenbeck 2003) run using the WAG substitution matrix and a gamma distribution with 4 rate categories and one category of invariable for 1,000,000 generations with sampling every 1,000 generations. Distance analyses were
performed using TREE-PUZZLE 5.2 (Schmidt et al. 2002) with four rate categories and 1 category of invariable sites. Tree-puzzle inferred alpha parameters were 0.89, 1.04 and 1.21 and the estimated proportion of invariable sites were 0.04, 0.05 and 0.10 for RPE, GAPDH and transketolase respectively. WEIGHBOR 1.0.1a was used to reconstruct distance trees (Bruno, Socci, and Halpern 2000). Bootstrapped distance matrices were generated using PUZZLEBOOT (shell script by A. Roger and M. Holder, http://www.tree-puzzle.de) with alpha parameter and proportion of invariable sites estimated using TREE-PUZZLE 5.2.

3.3 Results and Discussion

3.3.1 Ribulose-5-phosphate-3-epimerase

Ribulose-5-phosphate epimerase catalyzes the bidirectional conversion of ribulose-5-phosphate to xylulose-5-phosphate in both the Calvin cycle in the plastid of photosynthetic eukaryotes, and the cytosolic pentose-phosphate pathway of both photosynthetic and non-photosynthetic eukaryotes. Phototrophic eukaryotes therefore have two forms of this enzyme, and the plastid-targeted form in red algae, green algae and plants is cyanobacterial, whereas the cytosolic form is related to that of non-photosynthetic eukaryotes, as expected (Figure 3.1). The relationships between cytosolic epimerases of eukaryotes are poorly resolved, with only recently diverging groups such as vascular plants and metazoa recovered with strong bootstrap support.
Figure 3.1 Bayesian tree of RPE with PROML branch lengths

Tree was inferred using Mr. Bayes, with branch lengths estimated using PROML. Bootstrap values above 50% are shown. Values shown above node correspond to PHYML bootstrap support, those below node correspond to WEIGHBOR support. Eukaryotes are surrounded by grey shading. Filled circles are located beside Xylose-5-phosphate kinase/RPE gene fusions.
Previously, it has been shown that the plastid-targeted RPE of the chlorarachniophyte *Bigelowiella natans* is not related to other plastid-targeted or even cyanobacterial genes as one would expect, but is instead closely related to the gamma-proteobacterial genus *Pseudomonas* (Archibald et al. 2003). By increasing the sampling of RPE from other protist EST and genomic data, a similar proteobacterial RPE was found in five different chromalveolate genomes. Phylogenetic analysis confirmed that RPE genes from haptophytes *Emiliania huxleyi, Prymnesium parvum* and *Pavlova lutheri* are all of the gamma-proteobacterial type, as are those from the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (Figure 3.1). Relationships between the chromalveolate and *B. natans* genes are unresolved, but they collectively form a strongly supported group and the relationship with gamma-proteobacterial homologues, and more specifically the relationship with pseudomonads and alteromonads, remains well supported. The RPE from *B. natans* was reported to have a truncated N-terminal leader suggesting it may be plastid-targeted (Archibald et al. 2003). Further support for this prediction comes from both *P. tricornutum* and *E. huxleyi* genes, which have full-length leaders predicted to encode signal peptides at the N-terminus, which is consistent with the conclusion that this isoform of RPE is targeted to the plastid in these organisms as well. The ESTs encoding the RPEs of *P. parvum* and *P. lutheric* are N-terminally truncated, so it cannot be determined which compartment they are targeted to, though they do encode truncated leader sequences. Current models (v. 3.0) of the *Thalassiosira pseudonana* genome project predict that one of the RPE genes of *T. pseudonana* has an N-terminal extension not predicted to encode a signal peptide, the presence of a leader, or an initiator codon on the second *T. pseudonana* sequence has not been determined. Both diatom genomes encode another RPE that branch with the cytosolic homologues of other eukaryotes. In contrast, the cyanobacterium-derived, red algal RPE gene that would be expected to operate in the plastids of these organisms was sought in EST or genomic data from all six chromalveolates, but none was found. This includes the complete genome of *T. pseudonana* and the nearly complete genome of *P. tricornutum*, as well as the extensive EST databases for the three haptophytes. It would appear that the ancestral, cyanobacterium-derived RPE is most likely absent from these chromalveolates.
The origin and distribution of this gene in eukaryotes must be considered separately. The origin of the gene is addressed by the strong support for the eukaryotic genes being sister to a specific and taxonomically narrow group of bacteria, the pseudomonads. This argues for a relatively recent origin by lateral gene transfer from the pseudomonads to eukaryotes. The distribution, however, is more complicated since the eukaryotes that contain this class of RPE are not all closely related. There is evidence that haptophytes and diatoms are both members of the supergroup chromalveolates (Keeling et al. 2005), but chlorarachniophytes are members of a completely different supergroup, rhizaria. To explain this complex distribution by paralogy would be relatively simple if the enzyme were cytosolic: one would have to propose rhizaria and chromalveolates were closely related and that many lineages had lost the enzyme (considering only complete or nearly complete genomes, this would include apicomplexa, Perkinsus, and ciliates). However, the fact that this enzyme is targeted to the plastid significantly complicates this explanation because the *B. natans* plastid is derived from a green alga whereas the chromalveolate plastids are derived from a red alga. The plastid-targeted RPEs of both green algae and red algae are cyanobacterial (Figure 3.1) and complete genomes from neither group contain the proteobacterial type. Accordingly, for the plastid RPEs in chromalveolates and chlorarachniophytes to be derived from a common ancestor, the proteobacterial type would have to have coexisted with the cyanobacterial type in an ancestor of red and green algae followed by a complex pattern of reciprocal losses not only in rhizaria and chromalveolates, but also in red and green algae. If, alternatively, the enzyme was a cytosolic RPE in a hypothetical common ancestor of chlorarachniophytes, haptophytes, and diatoms, then it must have taken over plastid function twice independently, in addition to reciprocal losses. Both of these explanations are very complex and depend on higher order relationships among eukaryotes that are not known. In addition, the close specific relationship between the pseudomonads and these eukaryotes is more suggestive of a recent origin by lateral transfer than an ancient origin. Taken together, the simplest explanation for the current distribution is that RPE was transferred from a pseudomonad to some eukaryotic lineage and then transferred between two eukaryotic lineages (the direction cannot be inferred from the phylogeny since the topology of rhizarian and chromalveolate RPEs is not resolved).
On a side-note, we also observed an interesting gene fusion event involving the cytosolic RPE of the diatoms *T. pseudonana* and *P. tricornutum* and the prasinophyte green alga *Ostreococcus tauri*. In these three organisms, the cytosolic RPE is found as a xylulose kinase-RPE fusion-protein. In *P. tricornutum* and *O. tauri* the two proteins are part of an uninterrupted ORF, while in *T. pseudonana*, the xylulokinase and RPE are in different reading frames but it appears that this is most likely due to the presence of an intron. Xylulose kinase catalyses the phosphorylation of xylulose to xylulose-5-phosphate for entry into the pentose phosphate pathway. This reaction occurs immediately upstream of the ribulose-5-phosphate epimerase reaction, raising the intriguing possibility that the fusion protein may catalyze both reactions. Interestingly, the N-termini of *P. tricornutum* and *T. pseudonana* also encode a predicted signal peptide, suggesting that this fusion protein may be targeted to the plastid. *P. tricornutum* and *O. tauri* also encode canonical cytosolic RPEs related to other cytosolic isoforms. Gene fusion events involving carbon metabolic enzymes have been reported from other algae, such as those involving GAPDH and enolase in dinoflagellates (Takishita et al. 2005), and between triose phosphate isomerase and phosphoglycerate kinase in the mitochondria of heterokonts (Liaud et al. 2000). Whether the fusions arose in common or independently is not clear; *O. tauri* is a green alga while diatoms have plastids derived from red algae and the fusion genes are not demonstrably related (Figure 3.1), suggesting perhaps that the fusion arose twice independently.

### 3.3.2 Glyceraldehyde-3-Phosphate Dehydrogenase

Glyceraldehyde-3-phosphate dehydrogenase catalyzes the bi-directional conversion of glyceraldehyde-3-phosphate to 3-phosphoglycerate in both glycolysis and the Calvin cycle. GAPDH has been extensively sampled from eukaryotes and bacteria, revealing many cases of lateral transfer and paralogy. The typical eukaryotic cytosolic GAPDH is called GapC, whereas bacteria and plastids canonically use GapA/B. However, several eukaryotic groups also have GapA/B genes that are not derived from the plastid endosymbiont and are considered to have originated by lateral transfer (Markos, Miretsky, and Muller 1993; Viscogliosi and Muller 1998; Archibald et al. 2003; Takishita, Ishida, and Maruyama 2003). One of these cases is the divergent class of
GapA/B previously found only in diplonemids, heterotrophic relatives of kinetoplasts (Qian and Keeling 2001). Once again, however, with increased sampling we find the same class of GAPDH in haptophytes (*Isochrysis galbana*) and diatoms (*T. pseudonanna* and *P. tricornutum*). EST data from the haptophyte *E. huxleyi* also include three transcripts encoding a similar gene (GenBank accessions CX777351, CX776621, and EG034112), but they are too short to be included in the phylogeny. As described above, heterokonts and haptophytes are thought to be members of the same supergroup, chromalveolates, but they are not closely related to diplonemids, which are members of the excavates. Nevertheless, the GAPDH sequences from all three groups form a strongly supported clade, which in turn branches within a eubacterial group consisting of proteobacteria and cyanobacteria, as found previously for diplonemids alone (Qian and Keeling 2001). The chromalveolate sequences are paraphyletic in this analyses, as the diplonemids share a robust sister relationship with *I. galbana* to the exclusion of diatoms. The *I. galbana* GAPDH is not full length, but the N-termini of the two diatom sequences are of comparable length to GapA/B in bacteria and do not encode a predicted signal peptide, so these proteins are likely cytosolic.
Figure 3.2 Bayesian tree of GAPDH with PROML branch lengths

Tree was inferred using Mr. Bayes, with branch lengths estimated using PROML. Bootstrap values above 50% are shown. Values shown above node correspond to PHYML bootstrap support, those below node correspond to WEIGHtBOR support. Eukaryotes are surrounded by grey shading.
The distribution of this class of GAPDH bears many similarities to RPE: the gene is found in distantly related taxa (chromalveolates and diplonemids), each of which have relatives that lack it. Also in common with RPE, the close relationship between the eukaryotic genes and their bacterial sisters, together with their distant relationship to canonical eukaryotic GapC genes, indicates that the eukaryotic homologues originated by lateral transfer from a bacterium. However, this does not address how it came to its present distribution in eukaryotes. If there was a single transfer to the ancestor of chromalveolates and diplonemids, then the gene must have been lost in many of their relatives (considering only taxa where complete or nearly complete genomes are known, this includes apicomplexa, Perkinsus, ciliates, trypanosomatids, and perhaps Giardia and Trichomonas). However, given the large number of lateral transfer events already known to have involved GAPDH, including one between two eukaryotes (Takishita, Ishida, and Maruyama 2003), the current narrow range of taxa possessing this gene suggests instead that it was transferred from a bacterium more recently and subsequently spread to other eukaryotes by eukaryote to eukaryote transfers. The branching order between the eukaryotic genes is well supported, and taking this at face value suggests that there were either multiple transfers between eukaryotes or that the gene originated in chromalveolates and was transferred to diplonemids.
3.3.3 Transketolase

Transketolase, or glycoaldehyde-transferase catalyzes the reversible transfer of a ketol group between two 5 carbon sugars, producing a 3 carbon sugar and a 7 carbon sugar, or between a 4 carbon sugar and a 5 carbon sugar, producing a 6 carbon sugar and a 3 carbon sugar. Transketolase functions in the cytosol of non-photosynthetic eukaryotes, where it is involved in reactions of the classic pentose-phosphate pathway, and in the plastids of algae and plants, where it functions in the Calvin cycle as well as the reversible branch of the pentose-phosphate pathway. In at least some plants, an additional cytosolic isoform of transketolase, related to the plastid form exists (Bernacchia et al. 1995).

Transketolase exhibits a complex phylogenetic distribution across different groups of eukaryotes. Metazoa and ciliates have a highly divergent form of transketolase characterized by many gaps and deletions, and were not included in the alignment because they are difficult to align with other transketolases. Most eukaryotic cytosolic transketolases belong to a more conserved group that is widespread and form a single well supported clade (Figure 3.3). Similarly, most plastid-targeted transketolases have robust cyanobacterial affinities, as expected, although one with an unusually close relationship to the cytosolic clade described above. The phylogeny within the plastid-targeted clade is generally not well resolved, however, an interesting exception is the plastid-targeted transketolases from the euglenid Euglena gracilis and the dinoflagellate Heterocapsa triquetra, which form a very strongly supported branch in all analyses and which fall at the base of the plastid-targeted clade with strong support in analyses of the full protein (Figure 3.3). Euglenids and dinoflagellates are not closely related and their plastids are derived from green and red algae, respectively. These two sequences branching together is therefore unusual and reminiscent of a proposed transfer of a plastid-targeted GAPDH between euglenids and dinoflagellates (Takishita et al. 2005). The most interesting clade, however, comprises several eukaryotic transketolases that are unrelated to either the major cytosolic clade or plastid-targeted clade, but instead is closely related to the bacterial group, Chlamydiales (Figure 3.3). This clade includes not two, but several distantly related groups of eukaryotes. The chromalveolates are most
heavily represented, including the diatoms *T. pseudonana* and *P. tricornutum*, the haptophyte *I. galbana*, and the dinoflagellate with a haptophyte plastid *Karlodinium micrum*. There are also amoebozoans *Dictyostelium discoideum* and *Physarum polycephalum*, the excavate *E. gracilis*, and the rhizarian *B. natans*, amounting to four of the five supergroups (the exception being plants). In addition, more highly truncated ESTs were found from several other chromalveolates, specifically the haptophyte *P. parvum*, the cryptomonad *Guillardia theta* and another copy of the gene from *K. micrum*. These sequences were too short to include in the analysis shown in Figure 3.3, but in phylogenies restricted to the 3' end of the gene they consistently branched in this clade with high support (not shown). Amoebozoans and *E. gracilis* consistently form a strongly supported group, as do the chromalveolates and *B. natans*. Within the latter clade, *I. galbana* occupies a basal position in analyses based on the whole gene, and this is strongly supported by both bootstrap and by the presence of a unique conserved 4 amino acid insertion in *K. micrum*, *B. natans*, and the diatoms (Figure 3.4). It cannot be determined whether the transketolase genes of *B. natans*, *I. galbana* and *K. micrum* are cytosolic or plastid-targeted as they are N-terminally truncated. However, the two diatom diatom sequences are comparable in length at their N-termini to bacterial transketolases and cytosolic transketolases of other eukaryotes, and neither are predicted to encode signal peptides. Similarly, the *E. gracilis* transketolase is encodes a start codon in a position shared with the transketolase of *D. discoideum*, suggesting that it too may be cytosolic.

In addition to this group, there are also a few eukaryotic sequences that fall outside any of these transketolase groups. In particular, sequences from the amoebozoan, *Hartmanella vermiformis* and dinoflagellate *K. micrum* both have transketolases related to proteobacteria, planctomycetes and CFB bacterial clade, and that branch together with weak bootstrap support. (Figure 3.3). The *Entamoeba histolytica* transketolase, on the other hand, branches outside any eukaryotic clade but does not show an affinity to any bacterial group.
Figure 3.3 Bayesian tree of transketolase with PROML branch lengths.

Phylogenetic tree of transketolase. Tree was inferred from 473 amino acid characters using Mr. Bayes, with branch lengths estimated using PROML. Bootstrap values above 50% are shown. Values shown above node correspond to PHYML bootstrap support, those below node correspond to WEIGHBOR support. Eukaryotes are surrounded by grey boxes.
Figure 3.4 Boxshade alignment of a 4 amino acid indel common to diatoms, chlorarachniophytes and a dinoflagellate.

A nearby 11-15 amino acid indel characterizes a mixed group of CFB bacteria and proteobacteria as well as the amoebozoan *Hartmanella vermiformis* and the dinoflagellate *Karlodinium micrum*. Shared indels are surrounded by a box.
Distinguishing between lateral transfer and paralogy in transketolase is a more complex problem than in the RPE and GAPDH cases considered above, because the diversity of eukaryotes with the transketolase in question is much broader. At the same time, this breadth makes the transketolase case potentially much more interesting. This breadth makes a stronger case for paralogy – that this gene represents an ancient eukaryotic paralogue present in the last common ancestor of these groups. This implies the gene was lost in their close relatives, which is a substantial qualification because a large number of losses would be required. Considering only groups where genomes are complete or nearly complete, this would include animals, fungi, Entamoeba, apicomplexa, ciliates, Perkinsus, kinetoplastids, Giardia, and Trichomonas. The specific relationship to Chlamydiales is also difficult to reconcile with such an ancient origin, and suggest this gene may have originated more recently by lateral gene transfer from an ancestor of that group, but such an interpretation could suggest that no transfers between eukaryotes occurred if the gene was transferred to an ancient ancestor of most or all eukaryotes.

This strict interpretation runs into difficulties, however, when the phylogeny within the clade is considered. If no between-eukaryote transfer had occurred then the supergroups should be monophyletic or at least unresolved. This is not the case, since the rhizarian B. natans branches within the chromalveolates with strong support (Figure 3.3) and its relationship to K. micrum and the diatoms is further supported by the shared insertion (Figure 3.4). To explain these observations without lateral transfer it would be necessary to propose additional cases of paralogy since the gene originated in eukaryotes. The alternative is that the B. natans gene is derived from lateral transfer, which is consistent with the observation that nearly a dozen other B. natans genes are derived from other phototrophs by lateral transfer (Archibald et al. 2003). By extension, there is no reason to exclude the possibility that there have been several other transfers between eukaryotes (the present distribution could be achieved with as few as three transfers). This would also explain the punctate distribution without having to argue for loss in close relatives and is therefore the simplest explanation of the current data. If this is the
underlying mechanism by which this transketolase came to its present distribution, then it is second only to EFL in the complexity of its distribution.

3.4 Concluding Remarks

We describe three cases where the phylogeny of a carbon metabolic enzyme at first appeared to indicate a simple case of bacterium-to-eukaryote lateral gene transfer but with greater sampling showed this to be a more complex situation. In all three cases other eukaryotes with the same bacterial gene were discovered, and in each case these eukaryotes are distantly related to one another at the organismal level. The first important point to note here is that these observations were due to increasing the sample of eukaryotic molecular diversity, and therefore the distributions are likely to change again as sampling continues to grow. However, the distribution can only change in one direction – to greater complexity. It is possible that these genes will ultimately be found in such a large sample of eukaryotes that they will be concluded to represent ancient paralogues whose distribution is mostly due to gene loss, but given the high frequency of their absence in the current data, it seems more likely they will continue to be rare and punctate in distribution. In the case of the elongation factor-like GTPase further sampling has revealed more organisms that possess it (Gile, Patron, and Keeling 2006; Ruiz-Trillo et al. 2006), but it is still far less common than its counterpart, EF-1α.

A second noteworthy point is that we often contrast lateral gene transfer and lineage sorting as two contradictory possibilities, but may they be at work in parallel. If a new gene arrives in a lineage by lateral transfer, some descendants may keep it and some may lose it, resulting in an apparently complex distribution. Distinguishing this from serial transfers is difficult or perhaps impossible in certain circumstances, but we can still weigh the observations in favour of one possibility over the other. In particular, lineage sorting is probably more likely in cases of a complex distribution of presence and absence is found in closely related species, whereas serial transfers are more likely when the time frame is significantly longer. In the phylogenies described here, the time frame is very long indeed, and there is evidence from the internal phylogenies for eukaryote-eukaryote transfer. The significance of this lies beyond these three genes, in the process of transfer between eukaryotes in general. Currently there are few known cases of such transfers
because they are difficult to detect without better sampling than is currently available for most genes. The results reported here use an unusual feature of the gene, its origin from bacteria, as a flag to draw attention to subsequent transfers, but there is nothing to indicate the same process could not be occurring in many other genes where it is not as visible because they lack such flags.
3.5 Bibliography


Chapter 4 - Plastid-targeting peptides from the chlorarachniophyte *Bigelowiella natans.*

4.1 Introduction

Plastids, the light-harvesting organelles of plants and algae, are the product of an ancient symbiosis between a cyanobacterium and a non-photosynthetic eukaryote. This process is referred to as primary endosymbiosis, and has given rise to the plastids of green algae and land plants, red algae and glaucocystophytes. The primary plastids of red and green algae have also spread laterally amongst unrelated eukaryotes by a process called secondary endosymbiosis, in which a primary plastid-containing alga is engulfed and retained by a non-photosynthetic eukaryote (Archibald and Keeling 2002). Secondary plastid-containing organisms account for a significant fraction of present-day algal diversity: secondary algae are abundant, genetically diverse and contain a variety of plastid types. Secondary plastid containing lineages include the haptophytes, heterokonts, cryptomonads, dinoflagellates and apicomplexan parasites, which all contain red algal endosymbionts, as well as the euglenids and chlorarachniophytes, which contain green algal endosymbionts.

Chlorarachniophytes are unicellular, amoeboid flagellate algae found in marine environments that have acquired a plastid through secondary endosymbiotic uptake of a green alga (Hibberd and Norris 1984; Ludwig and Gibbs 1989; McFadden et al. 1994). As a consequence, the plastids of chlorarachniophytes are bounded by four membranes: the inner-two membranes are homologous to those of cyanobacteria and the primary plastids of green algae and plants, and the outer-two are derived from the plasma membrane of the green algal endosymbiont and the secondary host endomembrane system, respectively (McFadden 2001). As is the case in the red algal symbiont of cryptomonads, the chlorarachniophyte endosymbiont retains a highly reduced algal

nucleus, or "nucleomorph" (Hibberd and Norris 1984; Ludwig and Gibbs 1989; McFadden et al. 1994), nested between the second and third plastid membranes (i.e., in the residual cytosol of the endosymbiont).

The process of secondary endosymbiosis has important ramifications for protein targeting in secondary plastid-containing algae. This is because plastid genomes encode only a small fraction of the genes necessary to encode all plastid proteins. Most of these proteins are encoded by nuclear genes and the products are post-translationally targeted to the plastid using an amino-terminal extension called a transit peptide (McFadden 1999). Accordingly, most genes for plastid proteins in chlorarachniophytes would have been encoded in the nuclear genome of the green algal endosymbiont. However, preliminary sequencing of the nucleomorph genome shows that few such genes remain (Gilson and McFadden 1996): in the course of its secondary endosymbiotic integration, most of these genes were once again transferred, this time to the nuclear genome of the secondary host (Deane et al. 2000; Archibald et al. 2003). The protein products of these genes, therefore, must be targeted across four membranes to the plastid stroma (and in some instances across a fifth, thylakoid membrane). In all secondary plastids examined so far, this process involves the addition of a second N-terminal extension (for review see (McFadden 1999)). First, the protein is targeted to the host endomembrane system using a signal peptide, which directs the co-translational import of precursor proteins to the endoplasmic reticulum and is subsequently cleaved off (Blobel and Dobberstein 1975a; Blobel and Dobberstein 1975b). Second, precursor proteins are imported across the inner and outer chloroplast envelope following the general import pathway involving interaction of a transit-peptide with the plastid envelope and TOC and TIC complexes, common to the primary plastids of glaucocystophytes, red algae, green algae and plants (Schleiff and Soll 2000; Bruce 2001). How proteins are specifically directed to the plastid once they are in the host endomembrane system, and how proteins cross the second membrane from the outside (homologous to the algal cytoplasmic membrane, which has been lost in euglenids and dinoflagellates) are both unknown, and represent two of the outstanding mysteries in this process (McFadden 1999; van Dooren et al. 2001).

In cryptomonad, heterokont and haptophyte algae (chromists), the outer-most membrane of the plastids are continuous with the endoplasmic reticulum (ER) (Gibbs
1981; Ishida, Cavalier-Smith, and Green 2000), so plastid-targeted proteins are either co-
translationally inserted directly into the compartment where the plastid endosymbiont
resides (Gibbs 1979), or enter the ER and are translocated across the outer-membrane
through lumenal connections, as has been demonstrated in heterokonts with smooth outer
membranes (Ishida, Cavalier-Smith, and Green 2000). In contrast, the outer-most
membrane of the three membrane secondary plastids of dinoflagellates and euglenids are
not contiguous with the host endomembrane, and these organisms use an elaborate
system of vesicles to specifically target proteins to the plastid using the Golgi apparatus
(Sulli and Schwartzbach 1995; Sulli and Schwartzbach 1996; Sulli et al. 1999; Nassoury,
Cappadocia, and Morse 2003). Interestingly, these proteins are only partially translocated
across the ER membrane, so the mature peptide rests on the cytosolic face of the Golgi
vesicles. The outer-most membrane of chlorarachniophytes and apicomplexan plastids is
also not detectably contiguous with the host endomembrane, but the route taken by
plastid proteins in these organisms is not known, although it has been proposed that they
may travel through the Golgi (Bodyl 1997; Waller et al. 2000). Given that no other group
is known to direct proteins to their plastids through the lumen of the Golgi, this route may
be more difficult than previously conceived.

As a first step in characterising the route traveled by nuclear-encoded, plastid-
targeted proteins in chlorarachniophytes, we have analysed plastid-targeting leader
sequences from 45 plastid-targeted proteins from *Bigelowiella natans*, compared the N-
termini of these proteins to 10 ER-targeted and 38 cytosolic *B. natans* proteins, and tested
the heterologous function of one leader in the apicomplexan, *Toxoplasma gondii*. Overall
the characteristics of *B. natans* signal peptides and some characteristics of transit peptides
are similar to those found other well-studied systems, while other features of transit
peptides are different than those found in other algal groups.

4.2 Materials and Methods

4.2.1 Assembling a data-set of cytosolic, ER, and plastid proteins from *B. natans*.

Seventy-eight predicted plastid-targeted proteins were identified from a *B. natans*
EST project by (Archibald et al. 2003), and an additional gene encoding a full-length
plastid-targeted protein identified as a phosphoglycolate phosphatase precursor was added to this data-set. Of these 79 proteins, we have identified 45 clearly full-length transcripts with N-terminal extensions predicted to encode signal peptides (genbank accession numbers: AAO89070, AAP79136, AAP79140--AAP79142, AAP79144, AAP79147--AAP79150, AAP79152, AAP79153, AAP79155--AAP79161, AAP79164, AAP79166, AAP79167, AAP79170, AAP79174, AAP79175, AAP79177, AAP79179, AAP79181, AAP79183, AAP79187--AAP79189, AAP79192--AAP79196, AAP79199, AAP79203, AAP79208--AAP79211, AAP79216, AY611522.) cDNAs encoding putative plastid-targeted proteins were identified based on their phylogenetic relationship to plastid-targeted genes in other organisms, their homology with proteins involved in pathways specific to plastids, and their possession of a bipartite leader sequence consisting of a signal and transit peptide (Archibald et al. 2003). Three genes that encoded substantial N-terminal leaders but that lacked obvious signal peptides were excluded from further analysis. Ten ESTs for full-length genes encoding proteins known in other systems to reside within the endomembrane system or to be secreted were identified, and each of these was completely sequenced. Thirty-eight ESTs encoding full-length transcripts of cytosolic proteins were also identified in the same way, and each completely sequenced (table 4.1). The 49 newly sequenced genes were submitted to GenBank as accessions AY542966--AY543013, AY611522.
<table>
<thead>
<tr>
<th>Cytosolic and Nuclear</th>
<th>Endomembrane targeted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cofilin</td>
<td>Calreticulin</td>
</tr>
<tr>
<td>ADP ribosylation factor</td>
<td>Cathepsin Z</td>
</tr>
<tr>
<td>Histone H2B</td>
<td>COP protein</td>
</tr>
<tr>
<td>RAS related GTP binding protein</td>
<td>Cyclophilin B</td>
</tr>
<tr>
<td>Proteasome beta-subunit</td>
<td>Cysteine proteinase</td>
</tr>
<tr>
<td>Calmodulin</td>
<td>Digestive cysteine proteinase</td>
</tr>
<tr>
<td>Actin depolymerizing factor</td>
<td>Folate receptor homologue</td>
</tr>
<tr>
<td>Transcription factor BTF3</td>
<td>KDEL receptor</td>
</tr>
<tr>
<td>Small nuclear ribonucleoprotein SM-D1</td>
<td>Protein disulfide isomerase</td>
</tr>
<tr>
<td>Clathrin assembly protein</td>
<td>Serine threonine proteinase</td>
</tr>
<tr>
<td>Roadblock</td>
<td></td>
</tr>
<tr>
<td>Peroxidase</td>
<td></td>
</tr>
<tr>
<td>Glutathione-S-transferase</td>
<td></td>
</tr>
<tr>
<td>Actin related protein</td>
<td></td>
</tr>
<tr>
<td>Ubiquitin conjugating protein</td>
<td></td>
</tr>
<tr>
<td>Mago nashi homologue</td>
<td></td>
</tr>
<tr>
<td>Splicing factor</td>
<td></td>
</tr>
<tr>
<td>Calmodulin like myosin chain</td>
<td></td>
</tr>
<tr>
<td>RAB2</td>
<td></td>
</tr>
<tr>
<td>Small nuclear ribonucleoprotein</td>
<td></td>
</tr>
<tr>
<td>U6 snRNA associated SM like protein</td>
<td></td>
</tr>
<tr>
<td>F-actin capping protein beta subunit</td>
<td></td>
</tr>
<tr>
<td>Coronin</td>
<td></td>
</tr>
<tr>
<td>RAS-like GTPase</td>
<td></td>
</tr>
<tr>
<td>Ubiquitin conjugating enzyme E2-1</td>
<td></td>
</tr>
<tr>
<td>Ubiquitin conjugating enzyme E2-2</td>
<td></td>
</tr>
<tr>
<td>ADP ribosylation factor 1</td>
<td></td>
</tr>
<tr>
<td>Dynein 8 kDa light chain</td>
<td></td>
</tr>
<tr>
<td>RNA polymerase 2 15.9 kDa subunit</td>
<td></td>
</tr>
<tr>
<td>Selenoprotein W</td>
<td></td>
</tr>
<tr>
<td>Guanine nucleotide binding protein</td>
<td></td>
</tr>
<tr>
<td>Alpha tubulin 1</td>
<td></td>
</tr>
<tr>
<td>Alpha tubulin 2</td>
<td></td>
</tr>
<tr>
<td>Dynein outer arm light chain</td>
<td></td>
</tr>
<tr>
<td>U1 small nuclear ribonucleoprotein</td>
<td></td>
</tr>
<tr>
<td>Translation initiation factor 5A</td>
<td></td>
</tr>
<tr>
<td>Translation initiation factor 6</td>
<td></td>
</tr>
<tr>
<td>GTP binding nuclear protein RAN</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1 New cytosolic, nuclear and endomembrane-targeted proteins in *Bigelowiella natans*
4.2.2 Signal peptide predictions and analysis.

The N-termini of the plastid-targeted proteins were first analysed using the neural network prediction server SIGNALP v. 2.0 (Nielsen et al. 1997) to examine the first 50 amino acids for putative signal peptides. Three proteins were identified as having large N-terminal leaders that included a methionine codon, but were not predicted to encode a signal peptide and did not encode a stop codon upstream of the potential start codon. These proteins were excluded from further analysis due to the possibility that their leaders may not be completely sequenced. All signal and transit peptides were aligned according to their predicted cleavage sites. To verify these predictions and demonstrate examined according to several criteria. Perl scripts were written to generate sliding-window Kyte-Doolittle hydropathy profiles (Kyte and Doolittle 1982) and amino acid frequency profiles for the 15 amino acids upstream and 20 amino acids downstream of the predicted signal cleavage site using a five amino acid window. Amino acid frequencies were divided in to three categories representing properties known to be abundant or depleted in the signal and transit peptides in other organisms: hydroxylated (ST), basic (HKR) and acidic (DE) residues. The ten non-plastid, endomembrane-targeted proteins were analysed in the same way to serve as a positive control.

4.2.3 Transit peptide prediction and analysis.

Transit peptides were analyzed in a similar fashion to signal peptides, however transit peptide cleavage sites could not be assigned as reliably. To determine the approximate location of a potential cleavage site, the chloroplast transit peptide prediction server ChloroP (Emanuelsson, Nielsen, and von Heijne 1999) was used in combination with a multiple alignment of homologous proteins from other eukaryotes and eubacteria. In instances where ChloroP was unable to predict a transit peptide, or if the cleavage site of the transit peptide fell within the conserved portion of the mature sequence (as indicated by the alignment), or if the transit peptide was predicted to be much shorter than the remaining leader sequence, the end of the transit peptide was estimated based on the position in the alignment corresponding to the start of a cytosolic homologue in other eukaryotes or eubacteria. For this reason, putative transit peptide
cleavage sites represent only a rough estimate of the end of the transit peptide. Using the putative cleavage site as a point of reference for all transit peptides, 20 amino acids upstream and 20 amino acids downstream were analyzed for hydropathy and amino acid frequencies using the same approach as described for signal peptides. The overall amino acid composition of transit peptides was also analysed by concatenating the 45 inferred transit peptides and comparing their amino acid frequencies with those of 45 concatenated mature plastid-targeted proteins and the 38 concatenated cytosolic proteins.

4.2.4 Heterologous activity of *B. natans* plastid-targeting leader in *T. gondii*.

To test heterologous expression in *T. gondii* the coding sequence of *B. natans* ribulose 1,5-bisphosphate carboxylase (RuBisCO) was introduced into a parasite expression vector by recombination cloning. The sequence was amplified from cDNA clone by PCR using gene specific primers introducing half of the required attB recombination sites (5’-AAAAAGCGCTAAAAATGATGAGAAACGTTGCCCT 5’-AGAAAGCTGGGTACGAGGAGTAAGTGAATCCTCC) for 10 cycles. An aliquot of this reaction was used as template in a second reaction using and excess of universal attB primers and 5 low and 20 high stringency cycles (5’-GGGGACAAGTTTGTACAAAAAAGCAGGCT, 5’-GGGGACCTTTGTACAAAGAAGCTGGGT, see (Gubbels, Li, and Striepen 2003) for additional detail). The PCR product was cloned using a one tube BP/LR recombinase reaction using topoisomerase I treated destYFP/sagCAT destination vector (Striepen et al. 2002). The resulting plasmid RuBisCO-YFP, places the RuBisCO coding sequence under the control of the constitutive *T. gondii* alpha-tubulin promoter and in translational fusion upstream of yellow fluorescent protein.

RH-strain *T. gondii* tachyzoites were passaged in confluent human foreskin fibroblasts (HFF) and transfected essentially as described previously (Striepen et al. 2002). In brief, $10^7$ freshly harvested tachyzoites were resuspended in 300 µl cytomix and mixed with 50 µg plasmid DNA in 100 µl cytomix. Electroporation was performed using a BTX ECM 630 electroporator (Genetronics, San Diego, CA) set at 1500 kV, 25 Ω and 25 µF in a 2 mm cuvette. Parasites were inoculated into coverslip cultures and observed using a DM IRB inverted microscope (Leica, Wetzlar, Germany) equipped with a 100
Watt HBO lamp 24 hours after transfection. YFP and RFP expression was detected using appropriate filter sets (460/40 nm bp, em 527/30 nm bp, and ex 515/45 nm bp, em 590 nm lp respectively). Images were recorded using a digital cooled CCD camera (Hamamatsu, Bridgewater, NJ) and processed and analyzed using Openlab software (Improvision, Quincy, MA).

4.3 Results and Discussion

4.3.1 B. natans Signal Peptides

45 inferred plastid-targeted proteins from B. natans were predicted to encode N-terminal signal peptides when examined using SignalP. The predicted signal peptides varied in length from 16 – 47 residues, with a median length of 33 amino acids. Forty-four percent of the signal peptides examined (20 out of 45) possessed a motif consisting of a serine residue, followed by three neutral or hydrophobic amino acids, and an asparagine between 3 and 25 amino acids upstream of the cleavage site. Asparagines are uncommon in most signal peptides (Hoyt and Gierasch 1991) making the parallel occurrence of this motif unlikely. While the appearance of this motif in nearly half of the signal peptides examined is interesting, no role for this motif is obvious. This motif was absent in the signal peptides of predicted endomembrane resident and secreted proteins. The significance of this observation is unknown.

The predicted cleavage sites of B. natans signal peptides corresponded to a von-Heijne motif (von Heijne 1983; von Heijne 1984) with the –1 position occupied by an alanine residue in roughly 45% (21 out of 45) of the plastid targeted proteins examined, and by a glycine, serine or cysteine in all other cases. The –3 position is typically occupied by a variety of small uncharged amino acids such as alanine, valine, leucine, serine or threonine. The overall chemical properties surrounding the predicted cleavage sites of these 45 peptides were examined and compared with the equivalent region of proteins known to be targeted to the endomembrane system (Figure 4.1). In other systems that have been examined, signal peptides are generally rich in hydrophobic residues and small neutral residues, but depleted of acidic residues (Nielsen et al. 1997). The signal peptides of plastid-targeted proteins in B. natans conform to these general expectations:
alanine was the most abundant amino acid (not shown), while the acidic residues aspartic acid and glutamic acid were depleted (Figure 4.1, upper). These trends were also reflected in the signal peptides of *B. natans* ER-targeted proteins (Figure 4.1, lower).

Overall, the signal peptides of plastid-targeted proteins of *B. natans* appear to be chemically similar to those of secreted or endomembrane resident proteins, suggesting that plastid proteins are not likely distinguished by a specific type of signal peptide that directs plastid-targeted proteins to the appropriate compartment. Although signal peptides from some plastid-targeted proteins possess a distinct motif absent in the signal peptides of endomembrane resident or secreted proteins, no clear role for this motif is apparent.
Figure 4.1 Chemical properties of the signal/transit peptide boundary

Properties of the signal peptide–transit peptide boundary of 45 *Bigelowiella natans* plastid-targeted proteins (top) compared to the signal peptide–mature protein boundary of 10 ER-targeted proteins (bottom). X-axis corresponds to the position of the sliding window, while the Y-axis corresponds to Kyte-Doolittle values for hydrophobicity or the number of hydroxylated (ST), basic (HKR), or acidic (DE) amino acids per window divided by the size of the window respectively. All proteins are aligned on the predicted cleavage site, which is indicated by a dashed grey line.
4.3.2 *B. natans* transit peptides.

Transit peptides are typically more difficult to predict than signal peptides, but are generally variable in length, and plant transit peptides are typically rich in serine, threonine and arginine residues, and depleted in acidic amino acids (Keegstra, Olsen, and Theg 1989). The inferred transit peptides of *B. natans* were also highly variable in length, ranging between 20 and 83 amino acids. The overall amino acid frequency of these peptides was compared with that of mature plastid proteins and cytosolic proteins, and found to follow many of the expected trends (Figure 4.2). In particular, serine and arginine residues are enriched when compared with both mature plastid-targeted proteins and cytosolic proteins, while glutamic acid and aspartic acid are depleted. Interestingly, the positively charged amino acid lysine was the most depleted compared with both mature plastid-targeted proteins and cytosolic proteins, while this is the most abundant amino acid in the transit peptides of the malaria parasite, *Plasmodium falciparum* (Foth et al. 2003). This is interpreted as depletion since the frequency of lysine is lower in transit peptides than in both cytosolic and mature plastid proteins. In general, the mature plastid-targeted and cytosolic proteins shared similar amino acid frequencies, with the exception of alanine, which is more highly represented in cytosolic proteins (Figure 4.2). Altogether, the overall amino acid content of *B. natans* transit-peptides is within the bounds of expected properties, although the apparently extreme depletion of lysine does distinguish these sequences from other transit peptides.
Figure 4.2 Comparison of amino acid composition between transit peptides and mature and cytosolic peptides

Bars indicate the difference in percent amino acid composition between 45 concatenated transit peptides compared with the concatenated mature plastid proteins (grey bars) and 38 concatenated cytosolic proteins (black bars). Bars above the X-axis indicate an overabundance of that amino acid while bars below the X-axis indicate a depletion of that amino acid.
The chemical properties of the region surrounding the inferred transit peptide cleavage site were also analysed and compared with the N-termini of cytosolic proteins (Figure 4.3). In general, the region upstream of the cleavage site was slightly enriched in hydroxylated and basic residues (Figure 4.3, upper), a trend that is also noticeable at the signal-transit boundary (Figure 4.1, upper). Basic amino acids, in particular arginine, are concentrated near the predicted cleavage site of the transit peptide and less prominent near the N-terminus, which has also been reported for plant transit peptides (Claros, Brunak, and von Heijne 1997). Conversely, the frequency of acidic residues increases at the transit peptide-mature protein boundary, so that the overall basic nature of the transit peptides is more the result of a depletion of acidic residues than an overrepresentation of basic ones. These characteristics are also found in plant transit peptides, where they are thought to assist the plastid transit peptide in interacting with the negatively charged outer membrane of the plastid (Bruce 2001) as well as potentially serving as a charged binding site for processing peptidases (Richter and Lamppa 2002). These results are consistent with previous descriptions of the signal and transit peptides of the LHCII proteins in *B. natans* (Deane et al. 2000). In general, the transit peptides of *B. natans*, are enriched in the hydroxylated amino acid serine, the basic residue arginine, and depleted of acidic residues and lysine.
Figure 4.3 Chemical properties of transit/mature peptide boundary

Properties of the transit peptide–mature protein boundary of 45 plastid-targeted *Bigelowiella natans* proteins (top) compared with the amino terminal region of 38 cytosolic proteins. Axes and lines are as in Fig. 4.1.
4.3.3 Heterologous targeting in *T. gondii*.

Analyses of targeting peptide primary structure are often good guides for experimental design (Foth et al. 2003), but no transformation system exists for a chlorarachniophyte. Instead, we have tested one *B. natans* signal and transit peptide in the apicomplexan parasite *Toxoplasma gondii*. Because of their medical importance, the plastids of the apicomplexan intracellular parasites (apicoplasts) now rank among the best studied of any algal group, despite being the most recently discovered. Plastid targeting in *P. falciparum* and *T. gondii* have been examined in some detail, as have the characteristics of plastid-targeting leader sequences from *Plasmodium* (Waller et al. 1998; Roos et al. 1999; Yung and Lang-Unnasch 1999; DeRocher et al. 2000; Waller et al. 2000; Yung, Unnasch, and Lang-Unnasch 2001; Foth et al. 2003; Yung, Unnasch, and Lang-Unnasch 2003). In general, the apicomplexan signal peptides exhibit chemical characteristics typical of signal peptides in other organisms, but the transit peptides exhibit a number of unexpected properties. The transit peptides of *Plasmodium* in particular are highly enriched in lysine and asparagines residues (Foth et al. 2003). Currently, only four transit peptides have been described in detail from *Toxoplasma gondii*, these transit peptides are similar to plant transit peptides in being rich in hydroxylated and basic amino acids (DeRocher et al. 2000). Such characteristics are shared with the transit peptides of *B. natans*. The predicted *B. natans* RuBisCO transit peptide is similarly rich in hydroxylated residues, but with a lower frequency of arginines than that seen in other predicted targeting peptides.

Despite their divergent properties, the transit peptides of *Toxoplasma* and *Plasmodium* are reported to be interchangeable, and plant transit peptides are also reported to function in *Plasmodium* (Roos et al. 1999; Waller et al. 2000), although these observations need to be followed up with further experimentation. To test whether one *B. natans* signal and transit peptide possesses the biochemical characteristics necessary to function in a heterologous system featuring secondary plastids, transfection experiments in *T. gondii* were performed. This is not intended to be a comprehensive test of the interchangability of these sequences, but rather a preliminary test of our observations on the nature of *B. natans* peptides. Cells were transfected with a construct encoding YFP.
fused to the C-terminus of the leader from the *B. natans* RuBisCO small subunit protein. In the resulting transfectants (Figure 4.4B), YFP localizes to the parasitophorous vacuole, with some labeling of dense granules. When cotransfected with the secretory marker P30-RFP (Striepen et al. 2001), YFP and RFP co-localise (Figure 4.4D), indicating that the RuBisCO-YFP fusion protein is secreted, and not targeted to the plastid. Transfection of RuBisCO-YFP into a parasite line stably expressing plastid targeted FNR-RFP (Striepen et al. 2000) further confirms this observation, there is no overlap between FNR-RFP and RuBisCO-YFP labeling (Figure 4.4H). In the case of the RuBisCO leader, therefore, the predictions based on sequence characteristics are met, namely that the *B. natans* signal peptide is able to function in *T. gondii* while the chemical characteristics of the *B. natans* transit peptide are insufficient for targeting to the apicomplexan plastid. Though this may reflect the inability of *Toxoplasma* signal peptidase enzymes to recognize and cleave signal peptides specific to *B. natans*, thus preventing interactions between the transit peptide and apicoplast, it seems unlikely that such unprocessed proteins would be secreted in to the parasitophorous vacuole. Alternatively, the inability of this *Bigelowiella* transit peptide to function in *Toxoplasma* may represent a fundamental difference between the architecture of transit peptide recognition and plastid import in these two distantly related systems. Now that a significant number of *B. natans* plastid-targeting leaders are known, it would be interesting to test all such sequences for activity in apicomplexa to see if the RuBisCO example is representative of *B. natans* transit peptides in general, or if other chlorarachniophyte plastid-targeting sequences are sufficient to direct proteins in to the apicoplast. Further examination of complexes involved in protein translocation across the plastid envelope membranes in these two organisms may also shed some light on potential differences in plastid import. If the RuBisCO transit peptide is representative of typical *B. natans* plastid-targeting leaders, additional experiments would also be required to determine which characteristics are most critical to the function of *B. natans* targeting peptides, and which distinguish them from the targeting peptides of other organisms. The results presented here suggest that such features will more likely be specific to transit peptides than signal peptides, which appear to be universal amongst secondary plastid-containing algae.
Figure 4.4 Heterologous expression of the *Bigelowiella natans* RuBisCo small subunit leader peptide in the apicomplexan *Toxoplasma gondii*

Heterologous expression of the *Bigelowiella natans* RuBisCO small subunit leader peptide in the apicomplexan *Toxoplasma gondii*. A-D: co-expression of RuBisCO-YFP with secreted protein P30. Host cells were infected with *T. gondii* co-expressing RuBisCO-YFP fusion protein and P30-RFP or FNR-RFP fusion proteins respectively. (A) Phase contrast image of host cell (HC) showing a parasitophorous vacuole, the lumen of which (PVL) contains eight parasites (P). (B) RuBisCO-YFP fusion protein localizes to the parasitophorous vacuole (diffuse fluorescence around parasites) and dense granules (punctate fluorescence in parasites), both indicative of secretion. (C) The distribution of the established secretion marker P30-RFP is the same as that of RuBisCO-fusion protein, and the two show strong colocalization when the green and red channel are merged (D). This suggests that the RuBisCO leader directs the secretion of the protein. D-H: co-expression of RuBisCO-YFP with plastid-targeted protein FNR. (E) Phase contrast image of host cell and parasitophorous vacuole. (F) Expression of RuBisCO-YFP in same parasites. (G) The plastid in *T. gondii* is a single round organelle localized apical of the nucleus, which has been labeled by stable expression of FNR-RFP (Striepen et al. 2000). (H) Expression of RuBisCO-YFP in this background shows no colocalization (merged green and red channel).
4.4 Bibliography


Chapter 5 - The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts.

5.1 Introduction

Chlorarachniophytes are marine amoeboflagellates belonging to the recently recognized and diverse assemblage of protists called Phylum Cercozoa (Bhattacharya, Helmchen, and Melkonian 1995; Keeling 2001; Cavalier-Smith and Chao 2003a). Unlike the vast majority of cercozoans, chlorarachniophytes are photosynthetic, having acquired a plastid by secondary endosymbiosis of a green alga. Secondary endosymbiotic events have occurred on multiple occasions in the course of eukaryotic evolution and have involved hosts and endosymbionts from several different eukaryotic groups. Chromalveolates with plastids (cryptomonads, heterokonts, haptophytes, apicomplexa, and dinoflagellates) have secondary plastids derived from a red algal endosymbiont, and these have been hypothesized to trace back to a single endosymbiosis (Cavalier-Smith 1999; Fast et al. 2001; Patron, Rogers, and Keeling 2004). In contrast, euglenids and chlorarachniophytes have plastids derived from green algal endosymbionts (Gibbs 1978; Ludwig and Gibbs 1989; McFadden, Gilson, and Waller 1995; Van de Peer, Rensing, and Maier 1996); however, there is no clear indication of what kind of green alga gave rise to either plastid. Chlorarachniophyte and euglenid endosymbionts are most commonly believed to be derived from two independent endosymbiotic events (Delwiche 1999; Archibald and Keeling 2002), but they have also been hypothesized to have arisen from a single common secondary endosymbiosis, the so-called Cabozoa hypothesis (Cavalier-Smith 1999; Cavalier-Smith and Chao 2003b). The secondary endosymbiont of

chlorarachniophytes is also noteworthy because it has retained its nucleus and genome in a highly reduced form known as a nucleomorph (Gilson et al. 2006). The discovery of nucleomorphs, and the demonstration that they were degenerated algal nuclei, clinched the argument that plastids spread between eukaryotic lineages by secondary endosymbiosis (Whatley 1981), and they are still the source of important clues as to how secondary endosymbiosis works (Douglas et al. 2001; Cavalier-Smith 2002; Gilson and McFadden 2002). In addition to chlorarachniophytes, nucleomorphs are only found in one other group of algae, the cryptomonads, where they are derived from a red alga, (Ludwig and Gibbs 1987; Douglas et al. 1991). The photosynthetic organelle of chlorarachniophytes (and cryptomonads), therefore, contains two genomes, a highly reduced eukaryotic genome located within the periplastid space (between the two outer eukaryotic-derived membranes and the inner two membranes making up the plastid envelope), and a plastid genome within the stroma (McFadden et al. 1997). Proteins that function in the plastid of B. natans are, by extension, encoded in three separate genomes: the nucleus of the host (Deane et al. 2000; Archibald et al. 2003), the nucleomorph (Gilson et al. 2006), and the plastid itself. Ironically, of these genomes plastid proteins encoded in the nucleomorph and nucleus have been more intensively studied than those encoded in the plastid itself, and overall the least is known about the plastid genome in general.

Indeed, representative plastid genomes have been sequenced from all major groups of algae, except chlorarachniophytes. Complete genomes are known from at least one member of all three groups of primary plastids: glaucophytes, red algae and green algae (including plants and charophytes). Similarly, plastid genomes have been sequenced from secondary plastids of euglenids, cryptomonads, heterokonts, haptophytes, and apicomplexa. In addition, a great deal of data are known from the unusual genome of dinoflagellates, which is difficult to define as a genome because genes are encoded on single-gene mini-circles (Zhang, Green, and Cavalier-Smith 1999).

Here we describe the complete chloroplast genome from the model chlorarachniophyte B. natans. At 69.2 kbp, the B. natans chloroplast genome is the smallest chloroplast genome known from any photosynthetic eukaryote. Indeed, the B. natans plastid genome falls in the size range of plastid genomes from some non-
photosynthetic organisms. However, unlike *B. natans*, these genomes have lost a large number of genes relating to photosynthesis. The *B. natans* plastid genome has lost or transferred a few of the larger genes to the nucleus, but for the most part its reduced size is a result of compaction: small intergenic spaces, and the absence of introns. This genome also allowed us to carry out the first phylogenetic analysis with representatives of all major plastid groups and to test the Cabozoa hypothesis. A phylogeny of concatenated plastid proteins was conducted to test the relationship of the *B. natans* chloroplast to those of green algae, in particular euglenids. These analyses placed *B. natans* within the ulvophyte-trebouxiophyte-chlorophyte (UTC) group of green algae, at face value rejecting the Cabozoa hypothesis and supporting two independent origins for chlorarachniophyte and euglenid plastids.

5.2 Materials and Methods

5.2.1 Genome sequencing and annotation

Clones encoding chloroplast genomic DNA sequences were identified from the *B. natans* nucleomorph genome sequencing project (Gilson et al. 2006) by similarity to genes known to be plastid-encoded in most algae and plants. Assembly of these sequences resulted in 61 kbp of plastid sequence in 9 individual fragments. Gaps were filled by PCR amplification from one fragment end to all possible ends until a single, circular-mapping contig was acquired. All amplified fragments were cloned into pCR 2.1 vector by TOPO TA cloning (Invitrogen) and sequenced on both strands. Additional regions of ambiguous sequence were also amplified, cloned, and re-sequenced. One clone was found to contain a short, repeat-rich region resistant to sequencing in the intergenic region between *psbE* and *atpI* (a *trnaH* gene exists in the same intergenic space, but it was not part of the unsequenced region). The region in question was mapped by restriction digestion and determined to be approximately 100 bp in length. It was re-amplified and subcloned as progressively smaller fragments, but no additional sequence was obtained and we concluded the region likely has a highly stable structure making it difficult to sequence, and is too small to encode a gene.
All open reading frames larger than 100 bp were identified and their similarity to known genes determined by BLASTX searches (Altschul et al. 1990). RNA-encoding genes were sought by BLASTN searches. tRNA genes were identified using the tRNAscan online server (http://lowelab.ucsc.edu/tRNAscan-SE/). Since most of the genome consists of genes with a high degree of similarity to homologues in other green algal plastids, very few regions of ambiguous annotation remained, but all unassigned regions were searched by BLASTN and BLASTX.

5.2.2 Pulsed field gels

*B. natans* was grown in nutrient supplemented seawater (f/2) bubbled with filter sterilized air in continuous lighting at 24°C. The algae where harvested by centrifugation (3000g) and the cell pellet was resuspended in 10mM Tris-HCl, 100mM EDTA, 200mM NaCl and 0.5% molten low gelling temperature agarose at 37°C. The mixture was poured into a pre-chilled plug mold and once set the cell plugs were digested in 10mM Tris-HCl, 400mM EDTA, 1% N-lauryl sarkosyl and 1mg/ml Pronase E (Sigma) for 48 hrs at 50°C. The digested chromosome plugs were loaded into 1% agarose gels that were electrophoresed in a CHEF DRIII apparatus (BioRad ) in 0.5 x TBE at 14°C. To separate large chromosomes the pulse time was 100 s at 100 V for 3 hrs. This was then ramped over 36 hrs from 60-120 s at 200 V. Smaller chromosomes were separated with a pulse time of 20 s for 16 hrs at 175 V.

5.2.3 Phylogenetic analyses

Protein alignments were constructed for all protein-coding sequences identified in the *B. natans* genome using ClustalX (Thompson et al. 1997) and edited in MacClade 4.07. One exception was the ycf1 gene, which is highly divergent and proved to be too difficult to align for a meaningful analysis. Phylogenetic trees were generated for all alignments individually using PhyML 2.4.4 (Guindon and Gascuel 2003) with the Dayhoff substitution matrix and rates across sites modeled on a discrete gamma distribution with 8 variable site categories and one category of invariable sites. Concatenated datasets were analysed using PhyML 2.4.4 with the WAG substitution
matrix and site-to-site rate variation modeled on a discrete gamma distribution with 4 categories of variable sites and one category of invariable sites. Maximum likelihood analyses were also carried out using ProML 3.6 (Felsenstein 1993) with the JTT correction matrix and no gamma correction. Bootstraps for both methods were carried out in the same way. Bayesian analyses of the concatenated data was constructed using Mr. Bayes 3.0b4 (Ronquist and Huelsenbeck 2003) from 300,000 generations with sampling every 100 generations using the WAG substitution model, 4 gamma categories and one category of invariable sites. Branch lengths for Bayesian trees were inferred using ProML 3.6 with the JTT correction matrix and site rate variation modeled on a discrete gamma distribution with 4 rate categories with the alpha parameter and invariable sites obtained from PhyML 2.4.4 as above. All analyses were performed on the full data-set consisting of 56 proteins and 11,296 characters and a data-set with ribosomal proteins excluded resulting in 38 proteins and 9,108 characters. For the data-set with ribosomal proteins excluded, 50 burnin trees were removed from Bayesian analysis, whereas 60 were removed from the full data-set.

Approximately unbiased (AU) tests (Shimodaira 2002) were performed on the concatenated data-set excluding ribosomal proteins to compare several alternative positions of B. natans. Test trees were constructed by optimizing the phylogeny with B. natans excluded using Mr. Bayes 3.0b4 with same parameters as above (which resulted in an identical topology with the exception of B. natans being absent). B. natans was then added to 21 alternate positions, including as sister to E. gracilis. Site likelihoods for each tree were calculated using TREE-PUZZLE 5.2 (Schmidt et al. 2002) using the -wsl command with site-to-site rate variation modeled using the parameters from the original data-set. AU tests were carried out on site likelihoods using CONSEL 1.19 (Shimodaira and Hasegawa 2001).
5.3 Results and Discussion.

5.3.1 Genome structure

The *B. natans* chloroplast genome maps as a circle of 69,166 bp (Figure 5.1). This is consistent with results from pulsed field gel electrophoresis, from which the size is estimated to be ~70 kbp, and which also show the genome exists in complex concatenates (Figure 5.2). One small (100 bp) region between *psbE* and *atpI* could not be sequenced, but the sequence that was obtained from this intergenic region has several direct and inverted repeats, suggesting the possibility of a stable secondary structure. The overall GC composition of the genome is 30.2%, whereas coding sequences (protein-coding and RNAs) are 32.3% and non-coding is 16.1%, which is not unusual for a plastid genome. The genome includes inverted repeats of 9,380 bps comprising the SSU, LSU and 5S rRNA genes, several tRNAs, and genes encoding PsbM, PetD, PetB, and the large photosystem I apoprotein PsaA. With the exception of two protein-coding genes and 5 tRNA-encoding genes, the remainder of the genome is contained in the large single copy (LSC) region. The small single copy (SSC) region is reduced in comparison to that of plants and other algae, being only 4,124 bps (compared to typical SSCs which are between 10 and 20 kbp), and encoding only genes for PsaC and Ycf1. Many genes that are typically present in the SSC region are missing from the *B. natans* genome, including protein-coding genes such as *rpl32*, *cysT* and the NDH cluster.
Figure 5.1 Chloroplast genome of *Bigeloviella natans*

Genes on the outside are transcribed in the clockwise direction, and those on the inside are transcribed in the counter-clockwise direction. Genes are color-coded according to their function in photosynthesis (green), transcription/translation (red), or miscellaneous (purple). Transfer RNAs are indicated by their anti-codon and the amino acid they decode.
The chloroplast chromosome of *B. natans* occurs as 69 kbp concatamers. A. An ethidium bromide-stained gel of the chromosome-sized DNA molecules of *B. natans* electrophoresed beside the chromosomes of *Saccharomyces cerevisiae* (left). These were blotted and probed with the *B. natans rbcL* gene (right). The chloroplast chromosome migrates as linear 69 kbp concatamers that are probably the products of the breakage of chloroplast DNA circles. B. The sizes of the chloroplast DNA molecules as shown by Southern blot (right and indicated by arrows) can be seen in comparison to the ethidium bromide-stained mitochondrial (M) and nucleomorph chromosomes (I, II and III) when the pulsed field gel was run under different conditions.
Many unusual or unique rearrangements in gene order are also found in the *B. natans* chloroplast genome. Genes for photosystem proteins PsaA and PsaB are contiguous in all plastid genomes with the exception of *Chlamydomonas reinhardtii* and *Pseudendoclonium akinetum*, and this cluster has also been separated in *B. natans*. Similarly, an inversion has occurred in what would normally constitute the rRNA operon, so that the SSU and 5S rRNA genes are on the opposite strand from the LSU gene. The rRNA operon is a characteristic of nearly all genomes, but inversions breaking up the operon are a feature of only a few plastid genomes, for instance the apicomplexa (Wilson et al. 1996; Cai et al. 2003), zygnematales (Turmel, Otis, and Lemieux 2005), and the trebouxiophyte *Helicosporidium* (de Koning and Keeling 2006). In the ulvophyte *P. akinetum* the entire operon has also inverted so it is transcribed in the direction of the LSC region (Pombert et al. 2005).

A more unusual rearrangement has occurred in the major ribosomal protein operon. This cluster is conserved in many plastids and cyanobacteria, although losses have occurred in several lineages as well as lineage-specific fission and fusion events (Stoebe and Kowallik 1999). In most green algae and plants, a cluster of 12 genes between *rpl23* and *rpoA*, and a smaller cluster of *rps12*, *rps7* and *tufA* is all that remains of the original cyanobacterial operon. In *Euglena gracilis*, *rpoA* has been transferred to the nucleus and the genes for the remaining proteins have moved to other parts of the plastid genome, whereas in *C. reinhardtii* the operon ends at *rps8*. In *B. natans*, the operon has been split in a way similar to that seen in *C. reinhardtii*, except *rps8* has been translocated as well and the operon ends with *rpl5*. The gene order at the 3' end of the operon (*rps8* to *rpoA*) remains conserved, but has also been translocated.

Both the SSC and LSC exhibit some degree of strand bias that centres around the inversion of the rRNA genes. The rRNA genes are transcribed convergently, and so are the protein-coding genes flanking them in the SSC and many of the protein-coding genes in the LSC. In the right half of the LSC in Figure 5.1, all genes are transcribed towards the SSU rRNA, as are the block of genes proximal to the SSU rRNA on the left half of the LSC. The overall pattern of the genome has two points of divergence, one between *psbE* and *atpI* (the repeat-rich region that could not be sequenced) and another between *psaC* and *ycf1*, and two points of convergence between SSU and 5S rRNA. Strand bias
has been observed in other plastid genomes where genes tend to be transcribed away from the origin of replication, for example *E. gracilis* and *Helicosporidium* sp. (Hallick et al. 1993; de Koning and Keeling 2006), although other explanations seem to apply to other genomes (Cui et al. 2006). We have no direct evidence for a putative origin of replication in the *B. natans* plastid genome, but the strand bias and the existence of several direct and inverted repeats in the region between *psbE* and *atpI* (the unsequenceable region and also one of the two regions where transcription tends to diverge), all suggest this intergenic space is a good candidate.

5.3.2 Gene content, loss, and compaction

The *B. natans* plastid genome is considerably smaller than that of other photosynthetic eukaryotes, and marginally smaller than that of the non-photosynthetic parasitic plant *Epifagus virginiana* (Wolfe, Morden, and Palmer 1992). This reduction in size can be attributed to both gene loss and gene compaction (Figure 5.3). In terms of gene loss, the genome contains more genes than any non-photosynthetic plastid but fewer than any other photosynthetic plastid. We identified 57 protein-coding genes, 4 of which are duplicated in the inverted repeat (giving a total of 61). While this is less than any of the photosynthetic plastids, it is comparable to the 66 genes in the much larger genome of *E. gracilis*, or the 69 in *C. reinhardtii*, the largest completely sequenced chloroplast genome. Overall, there has been some gene loss in the *B. natans* genome, but not much: compared to all other publicly available green algal genomes, 8 genes common to all these algae are absent in *B. natans*. When this comparison is expanded to include *E. gracilis*, only 3 genes common to this group are absent in *B. natans*, and if this is further expanded to include photosynthetic plant genomes in this set, only one loss is unique to *B. natans*, *psbZ*. All genes encoding photosystem proteins found in any other green algae have been retained, with the exception of *psal*, *psaM*, and *psbZ*. Additionally, all of the cytochrome components found in other green algae with the exception of *petL* have been retained, as have all of the ATP synthase genes found in other green algae.
Figure 5.3 Histogram of plastid genome sizes

Histogram representing plastid genome size and coding capacity ranked by coding capacity. Bars are divided into protein- and RNA-coding sequence (black, bottom), intron content (grey, middle) and intergenic spacers (white, top). Genomes are ranked according to the amount of coding DNA and ordered from left to right by increasing amount of coding sequence.
The complement of ribosomal protein genes is also similar to that of green algae. *rpl12, rpl32* and *rps9* are absent, but none of these losses are unique among green algae or streptophytes. Nucleus-encoded genes for plastid-targeted *rpl12* and *rps9* have already been reported from *B. natans* (Archibald et al. 2003). Interestingly, these two genes are contiguous in other green algae with the exception of *C. reinhardtii* and *Mesostigma viride*, raising the possibility that these two genes may have been transferred to the host nucleus together.

Much of the reduction in gene content in *B. natans* comes from genes with unidentified function (*ycf* genes) and genes with miscellaneous functions in chlorophyll biosynthesis, (*chlB, chlL, chlN*), cytochrome biogenesis (*ccsA*), fatty acid metabolism (*accD*) and cell division (*ftsI, ftsH, ftsW, minD, minE*). Also absent are all genes for NDH proteins, which are absent from green algae with the exception of *Nephroselmis olivacea*. Many of these genes are often found in the SSC region, which is considerably reduced in *B. natans*. Parallel gene loss has been shown to be relatively common in plastid genome evolution (Martin et al. 1998). We plotted gene losses on tree topologies inferred by Bayesian and likelihood methods (see below), and found that 23 losses are predicted to have occurred in the lineage leading to *B. natans* since it diverged from its last common chlorophyte ancestor (Figure 5.4). This is more than most lineages, but comparable to *E. gracilis* and *C. reinhardtii*.

As alluded to earlier, gene loss only partly accounts for the small size of the *B. natans* plastid; the *B. natans* plastid genome is also unusually gene-dense. To illustrate this point, compare the *B. natans* genome with those of *E. gracilis* and *C. reinhardtii*, both of which encode similar numbers of genes but are much larger (Figure 5.3). Whereas the *E. gracilis* genome is characterized by large numbers of introns, the *B. natans* genome contains no introns whatsoever, not even the typically conserved tRNA*Leu* intron also found in cyanobacteria (Kuhsel, Strickland, and Palmer 1990). The *C. reinhardtii* genome has an average intron content, but has large intergenic spaces. In contrast, the intergenic spaces in the *B. natans* genome are severely reduced. Average intergenic space is only 91 bp, which is comparable to the apicoplast genomes of parasites such as *Theileria parva* and the large gene-rich plastids of red algae and the heterokont *Odontella sinensis*. 
Figure 5.4 Plastid gene loss events plotted on the phylogeny of concatenated plastid genes.

Plastid gene loss events plotted on the phylogeny based on concatenated plastid genes. Unique gene loss events are indicated in blue font, gene gain events are indicated in red font.
5.3.3 tRNA genes

The *B. natans* plastid genome encodes 27 tRNAs. One species of tRNA is found for each amino acid, except for serine and glycine, which have two each, and leucine and methionine, which have three each. The three methionyl tRNAs correspond to the initiator-tRNA (f-Met), the elongation methionyl-tRNA, and the modified isoleucyl-tRNA. Intriguingly, the euglenids *E. gracilis* and *E. longa* possess the exact same complement of tRNAs. With wobble rules considered, this complement of tRNAs is near but not exactly the minimum set of tRNAs (de Koning and Keeling 2006), so why do they share the same set? Although the tRNA content between *B. natans* and euglenids is the same, chloroplasts genomes have descended from an already limited subset of tRNAs, so such convergence may not be unlikely when considering the limited subset of tRNAs present in all plastid genomes.

5.3.4 Phylogenetic relationship to other plastid genomes

To investigate the origin of the *B. natans* endosymbiont, we have inferred phylogenetic trees from concatenated data-sets of nearly all protein-coding genes in the chloroplast genome, as well as individual gene phylogenies for each protein-coding gene. Individual phylogenies were reconstructed for 56 of the 57 protein-coding genes in the *B. natans* plastid genome. One protein, Ycf1 was not included in the analysis as it proved too divergent and difficult to align. Overall, most of the individual phylogenies place *B. natans* within the Chlorophyta with good support, but without any consensus as to which group of green algae is sister to *B. natans* (not shown).

Analyses of concatenated genes were also carried out. Several previous studies have used concatenated plastid proteins to address a variety of questions, and one issue that has emerged is the divergent nature of the ribosomal proteins and their potentially misleading contribution to the phylogeny. This was recently shown relating to the monophyly of the chromists (Hagopian et al. 2004). We have accordingly inferred phylogenies using both the full set of 56 proteins (11,296 characters) and a slightly reduced set excluding the ribosomal proteins (38 proteins and 9,108 characters). The tree
of concatenated proteins excluding ribosomal proteins is shown in Figure 5.5. Overall the tree resembles other analyses of similar data (Martin et al. 2002; Matsuzaki et al. 2004; Hagopian et al. 2004), with well-supported groups for the red plastid lineage (with a monophyletic and well-supported chromist subgroup), and distinct streptophyte and chlorophyte groups. The glaucophyte *Cyanophora paradoxa* branches as sister to green algae and plants, a topology which has been recovered in similar analyses with ribosomal proteins excluded (Hagopian et al. 2004). *B. natans* branches definitively within the Chlorophyta, and more specifically within the clade consisting of ulvophytes, trebouxiophytes, and chlorophytes (collectively the UTC group), although the position of *B. natans* with regard to specific members of the UTC clade is equivocal. The branching order within the UTC has previously been shown to differ between analyses: in Figure 5.5 the chlorophyte *C. reinhardtii* branches first, in accordance with recent analyses based on concatenated chloroplast-encoded genes from green algae (Pombert et al., 2005). Significantly, *E. gracilis* was never observed to branch within the well-supported UTC/chlorarachniophyte clade.

Analyses using the entire 56 gene data-set were also performed, and no difference was found in most of the well-supported branches, with the exception of chromists, which did not emerge as a monophyletic clade using the full data-set (Figure 5.6). The UTC clade including *B. natans* was recovered with similar bootstrap support. Because plastid genomes encode slightly different repertoires of proteins, in particular *B. natans* and *E. gracilis*, we also constructed PHYML trees from concatenated data-sets with all gaps removed. These trees were based on 9,107 characters, and they produced similar topologies. In particular, PHYML support for the UTC clade including *B. natans* remained 98% (not shown).

Approximately unbiased (AU) tests were performed on the data-set excluding ribosomal proteins to compare several different positions of *B. natans*, including a sister relationship to *E. gracilis* (i.e., the Cabozoa hypothesis) and a basal relationship to all Chlorophyta. All alternative topologies were rejected at the 5% percent confidence level, except two topologies, a sister relationship between *B. natans* and the entire UTC clade and a sister relationship between *B. natans* and chlorophytes.
Figure 5.5 Protein maximum likelihood tree of concatenated plastid-encoded genes.

The tree was constructed from 38 proteins amounting to 9,108 amino acid positions (complete set excluding ribosomal proteins). The tree topology was inferred using Bayesian analysis with maximum likelihood branch lengths. Numbers at nodes correspond to bootstrap support from ProML (top), PhyML (bottom). Distance analyses carried out on the same alignment with missing data removed recovered the B. natans plus UTC clade with 100% support. Filled circles correspond to alternate topologies that failed AU tests at a 5% confidence level, open circles indicate topologies that cannot be rejected at a 5% confidence level.
Figure 5.6 Protein maximum likelihood tree of concatenated plastid-encoded genes

The tree was constructed from 56 proteins and 11,296 amino acid positions (complete set including ribosomal proteins). Numbers at nodes correspond to bootstrap support from ProML (top), PhyML (bottom). Distance analyses carried out on the same alignment with missing data removed recovered the B. natans plus UTC clade with 100% support. The tree was inferred as in Figure 5.5.
5.3.5 Origin of chlorarachniophyte plastids

With the aim of explaining plastid diversity with as few endosymbiotic events as possible, the Cabozoa hypothesis suggested that the green algal endosymbionts of chlorarachniophytes and euglenids shared a common origin (Cavalier-Smith 1999; Cavalier-Smith 2003). Chlorarachniophytes and euglenids are thought to belong to two different super-groups of eukaryotes that are principally non-photosynthetic; the chlorarachniophytes to the Rhizaria and the euglenids to the Excavata (see Keeling et al. 2005 for review). Excavates include a diversity of non-photosynthetic groups like diplomonads, retortamonads, parabasalids, oxymonads, and jakobids. Euglenids are the only photosynthetic excavates and are known to be specifically related to a subgroup of non-photosynthetic excavates, kinetoplastids and diplonemids. Rhizaria comprises foraminiferans, cercozoans, and some radiolarians and heliozoans. Like excavates, rhizaria are primarily non-photosynthetic. Chlorarachniophytes are the only rhizaria known to have secondary endosymbionts, though the thecate filose amoeba *Paulinella chromatophora* has a cyanobacteria-derived photosynthetic organelle unrelated to the primary plastids of other eukaryotes (Marin, Nowack, and Melkonian 2005). Like euglenids, chlorarachniophytes are derived cercozoans, recent phylogenies of the cercozoa suggest they are a sister group to filosa (Bass et al. 2005). Taken to its necessary conclusion, the Cabozoa hypothesis predicts that excavates and rhizaria share a common photosynthetic ancestor, and therefore that the majority of both excavates and rhizarians have lost photosynthesis.

The improbability of these multiple losses of photosynthesis is, in the Cabozoa hypothesis, counterbalanced by the improbability of secondary symbioses occurring twice, given the difficulties implicit in the *de novo* evolution of targeting machinery in independent lineages (Cavalier-Smith 1999). This is a difficult argument to sustain because we have no appreciation of the relative probabilities of these two events. Indeed, plastid loss is arguably more difficult than gain, since an organism could become dependent on non-photosynthetic metabolic pathways such as fatty acid, isoprenoid and heme biosynthesis that plastids can bring with them. Even making the important distinction between 'plastid loss' and 'photosynthetic loss', the Cabozoa hypothesis
demands many plastid loss events in organisms with complete or near complete genome sequences from which no plastid data are in evidence (e.g. trypanosomes, trichomonads, diplomonads).

The Cabozoa hypothesis makes no predictions about what kind of green alga gave rise to the chlorarachniophyte and euglenid endosymbionts, it does require that they are related to the exclusion of other green algae. It is of course possible that rhizarians and excavates do share a common ancestor (there are currently no data supporting or refuting this), but the Cabozoa hypothesis also requires that their common ancestor already had a plastid. Therefore, plastid sequence data can potentially disprove the Cabozoa hypothesis by showing one or both of chlorarachniophyte or euglenid plastids is more closely related to any other green algal plastid than they are to one another. Our concatenated analyses support a close relationship between *B. natans* and UTC green algal plastids to the exclusion of *E. gracilis*, arguing against a single secondary endosymbiosis of green plastids and the Cabozoa hypothesis.

5.4 Conclusions.

The chloroplast genome of *B. natans* is the first chloroplast genome from a chlorarachniophyte, the last major algal lineage for which a chloroplast genome has not been sequenced. It is also the smallest chloroplast genome known to date from a photosynthetic eukaryote, although it encodes most of the genes found in other photosynthetic green algae and plants. Chloroplast genomes of photosynthetic green algae display a large variation in size (Simpson and Stern 2002), those completely sequenced range between 150 and 200 kbp, but this may be only a small subset of the diversity that exists. Restriction digest suggest that the chloroplast genome of the ulvophyte, *Acetabularia mediterranea* is larger than 400 kbp (Tymms and Schweiger 1985), and physical maps of the plastid genome of the ulvophyte *Codium fragilis* suggest that it is only 89 kbp (Manhart et al. 1989). Although the genome of *B. natans* is smaller than any chloroplast genome yet reported, it is possible that the discrepancy in size between the genome of *B. natans* and that of other green algae may not be so dramatic and that *B. natans* simply lies at the lower end of a diverse spectrum of algal chloroplast genome sizes.
The origin of the chlorarachniophyte endosymbiont has been a topic of controversy since its discovery. Pigment composition was used to suggest a prasinophyte origin of the endosymbiont (Sasa et al. 1992). In contrast, molecular data has suggested a trebouxiophyte (Van de Peer, Rensing, and Maier 1996), and more recently an ulvophyte origin of the *B. natans* endosymbiont. (Ishida et al. 1997; Ishida, Green, and Cavalier-Smith 1999). Our analyses do not distinguish between an ulvophyte, trebouxiophyte or chlorophyte origin for the endosymbiont, but they do preclude a prasinophyte, streptophyte or deeper chlorophyte origin of the chlorarachniophyte plastid. Similarly, we recover no support for a clade of chlorarachniophytes and euglenids, arguing against the Cabozoa hypothesis. Taken together, our data suggests that the plastids of chlorarachniophytes are related to a derived group of green algae, and that the plastids of euglenids and chlorarachniophytes are of distinct and independent origin.
5.5 Bibliography


de Koning, A. P., and P. J. Keeling. 2006. The complete plastid genome sequence of the parasitic green alga, *Helicosporidium* sp. is highly reduced and structured. BMC Biol. 4:12.


Chapter 6 - Conclusions

6.1 Summary

The preceding chapters have dealt with various related topics in chlorarachniophyte evolution. Chapter two, details three possible endosymbiotic gene transfers of three Calvin cycle genes from the green algal endosymbiont of chlorarachniophytes to their host nucleus and subsequent re-targeting of these genes to the plastid. The three enzymes discussed in this chapter are among the few known Calvin cycle enzymes in chlorarachniophytes that show evidence of being derived from the green algal endosymbiont rather than through lateral transfer of a gene encoding this enzyme from another organism. Chapter three discusses two other Calvin cycle enzymes and an additional pentose phosphate enzyme. These enzymes belong to a set of metabolic enzymes in chlorarachniophytes that show signs of having been acquired from other eubacteria or eukaryotes through lateral gene transfer (Archibald et al. 2003). Chapter four describes the bipartite targeting leaders of chlorarachniophytes and their comparison to targeting sequences from other organisms. Examination of plastid-targeting leaders is an essential supplement to the study of endosymbiotic gene transfer, as a transferred gene is often re-targeted to the plastid. Finally, chapter five characterises the plastid genome of *Bigelowiella natans*. The plastid genome is the original source of all endosymbiotic gene transfers, accordingly an essential component of understanding endosymbiotic gene transfer is organellar genomics. Below is a brief recapitulation of each chapter with the importance of the findings and future directions of each chapter outlined.

6.2 Transfer and recompartmentalisation of carbon metabolic genes

Chapter two describes the two fructose-bisphosphate aldolase genes of *B. natans* in addition to touching on several gene replacement and recompartmentalisation events in the history of photosynthetic organisms. The original plastid-targeted eukaryotic aldolase was likely a cyanobacterial class II aldolase. This ancestral condition seems to be preserved in the glaucocystophytes *Cyanophora paradoxa* (Gross et al. 1999; Nickol et al. 2000). Glaucocystophytes are by many accounts the earliest branch of primary plastid
bearing eukaryotes (Cavalier-Smith 1982), making the possession of this class II aldolase a possible gene character for rooting the tree of photosynthetic eukaryotes. Chlorarachniophytes have both plastid and cytosolic isoforms of class I aldolase. The plastid class I aldolase of chlorarachniophytes may be derived from an endosymbiotic gene transfer event. Chlorarachniophytes and euglenids both have an additional cytosolic class II aldolase with obscure eukaryotic affinities. Finally, diatoms have a plastid-targeted type A class II aldolase that is clearly distinct from the type B class II aldolase of glaucocystophytes. Patron et al. (2004) expand on this clade of diatom aldolases by including plastid-targeted class II aldolases from haptophytes, cryptomonads and dinoflagellates in addition to cytosolic forms of aldolase from these same groups. The possession of a class II type A plastid-targeted aldolase was argued by Patron et al. (2004) to be a distinguishing molecular character of photosynthetic members of the chromalveolates. The acquisition of this plastid-targeted aldolase could have occurred either through the duplication of a pre-existing cytosolic aldolase, or through the re-targeting of a class II red algal aldolase to the plastid.

The plastid-targeted sedoheptulose-bisphosphatase and fructose-bisphosphatase enzymes of chlorarachniophytes are possible products of endosymbiotic gene replacement, though the case is not clear-cut for either enzyme due to the basal position of SBPase and plastid FBPase. My description of SBPase like enzymes from fungi, and the clustering of these fungal SBPases with a described SBPase from Trypanosoma brucei casts doubt on the proposed origin of the kinetoplastid SBPase through endosymbiotic gene transfer from an ancestral kinetoplastid plastid (Hannaert et al. 2003). Though, as noted in chapter 2, no cyanobacterial homologues of SBPase exist, making the endosymbiotic origin of even plant and algal SBPase uncertain. Class I aldolase phylogeny also fails to group T. brucei with plants, refuting earlier analyses that suggested that this enzyme as well may have been acquired from a photosynthetic organelle (Hannaert et al. 2003). As SBPase is a plastid-targeted enzyme, its existence in non-photosynthetic organisms remains a mystery, though it has been suggested it might function in the pentose phosphate pathway in Trypanosoma (Hannaert et al. 2003). A lateral transfer event from eubacteria to Toxoplasma gondii was additionally described in
chapter two, a eubacterial form of FBPase clearly unlike the eukaryote related FBPase of other eukaryotes was reported from this coccidian.

6.3 Evidence for inter and intra domain transfer

A recent survey of ESTs encoding plastid-targeted proteins from the chlorarachniophyte *Bigelowiella natans* revealed a surprising number of lateral gene transfer events in the history of this organism (Archibald et al. 2003). A large number of these transfers involved genes encoding metabolic enzymes functioning in the Calvin cycle. The genes encoding glyceraldehyde-3-phosphate dehydrogenase and ribulose-5-phosphate-epimerase in *B. natans* were shown to have a gamma-proteobacterial ancestry. Others, such as the gene encoding phosphoglycerate kinase grouped *B. natans* with phototrophic eukaryotes unrelated to the chlorarachniophyte endosymbiont. The small subunit gene of RuBisCO (*rbcS*) grouped *B. natans* with streptophytes and was shown to encode an indel shared with this group. Some genes that initially produced unresolved phylogenies, or suggested endosymbiotic gene transfers have since been more extensively sampled and re-analysed and show relationships to other eukaryotes as well. Such has recently been the case for the Calvin cycle enzyme phosphoribulokinase (Petersen et al. 2006). Of all the Calvin cycle enzymes identified in *B. natans*, only three, fructose bisphosphatase, fructose bisphosphate aldolase I, and sedoheptulose-1,7-bisphosphatase show possible relationships to green algae. Of these three, fructose bisphosphate aldolase I and fructose bisphosphatase show only weak relationships to green algae, and sedoheptulose bisphosphatase branches at the base of green algae and plants. (Rogers and Keeling 2003).

It has been suggested that as heterotrophs inhabiting environments rich in prokaryotes and other phototrophic eukaryotes, chlorarachniophytes may be uniquely equipped for obtaining new genes through lateral gene transfer (Archibald et al. 2003). Genes acquired from other phototrophic eukaryotes might have an advantage as potential plastid-targeted proteins as a result pre-existing transit peptides, and depending on the donor of the gene, signal peptide. Genes acquired from prokaryotes, such as GAPDH and RPE would have to acquire targeting presequences *de novo* (Archibald et al. 2003). We have set out to analyze two of the phylogenies of Calvin cycle enzymes previously
reported as lateral gene transfers in *B. natans* to determine if these transfer events are unique to *B. natans* or if these same genes have been transferred from prokaryotes to eukaryotes in other lineages of eukaryotes.

My analyses expand on the gamma-proteobacterial clade of RPE, including *B. natans*, as well as a bacterial GAPDH clade containing the GAPDHs of *B. natans* and diplonemids. Re-construction of the RPE tree reveals that chromists have a similar prokaryote-related enzyme to that of *B. natans*. The observation that the RPEs of at least some chromists and chlorarachniophytes both appear to be plastid-targeted makes an argument for an ancient transfer or paralogy untenable for reasons discussed in chapter 3. Instead, I invoke a lateral transfer event between chlorarachniophytes and chromists following the acquisition of a eubacterial gene in one of these lineages. Re-construction of a GAPDH phylogeny with GAP A/B sequences from chromistans suggests that the acquisition of a proteobacterial GAPDH by *B. natans* was an independent event, though the transferred GAPDHs of diplonemids and chromists branch together. Given the extensive history of lateral transfers of GAPDH among eubacteria (Figge and Cerff 2001), between eubacteria and eukaryotes (Markos, Miretsky, and Muller 1993; Wiemer et al. 1995; Qian and Keeling 2001; Archibald et al. 2003), eukaryotes and eukaryotes (Takishita, Ishida, and Maruyama 2003) and possibly even from eukaryotes to eubacteria (Figge et al. 1999), it is plausible that another eukaryote to eukaryote gene transfer has occurred between diplonemids and chromistans subsequent to the acquisition of a prokaryotic GAPDH by one of these lineages. This conclusion is further supported by the absence of this diplonemid like gene in other euglenozoans, suggesting that the diplonemid GAPDH was not present in the common ancestor of diplonemids and kinetoplastids, but rather acquired recently from a prokaryote-eukaryote or eukaryote-eukaryote lateral transfer event. Similarly, this same gene is absent in relatives of chromistans, including complete genomes of apicomplexans and ciliates. Transketolase produces the most confusing topology of all three analyses, grouping cyanobacteria with non-photosynthetic eukaryotes and producing a scattered assemblage of eukaryotes across the tree. One clade of eukaryotes with affinities for the transketolase of chlamydiales includes a diversity of organisms from four different eukaryotic supergroups, excavates, rhizaria, opisthokonts and chromalveolates. I suggest that given the
disparate distribution of this enzyme across different eukaryotic supergroups, a shared in
del between chlorarachniophytes and chromistan transketolases and the surprisingly strong support for grouping *B. natans* with chromists, that a lateral transfer event involving the transketolase of *B. natans* and a chromistan has occurred. Further lateral transfer events between opisthokonts, euglenids and chromists may further account for the presence of this bacterial type transketolase in eukaryotes.

**6.4 Plastid-targeting leaders**

Prior to the paper included as chapter 4, a large-scale study of chlorarachniophyte targeting peptides had not been conducted. Archibald et al. (2003) identified ESTs encoding plastid-targeted proteins on the basis of signalP predictions, the presence of N-terminal leaders and homology to other plastid proteins. An earlier study by Deane et al. (Deane et al. 2000) identified plastid-targeted LHC proteins and described their bi-partite targeting signals as consisting of an N-terminal hydrophobic signal peptide followed by a length of peptides rich in serine residues. Examination of 45 plastid-targeting leaders from *B. natans* reveals similarities between the signal peptides of plastid-targeted peptides and secreted peptides in *B. natans*, both are hydrophobic and deficient in acidic residues. The characteristics of stromal targeted peptides are similar to those described by Deane (Deane et al. 2000) overall, and suggests that in many ways the stromal targeting peptides of chlorarachniophytes are similar to those described from plants in being rich in hydroxylated residues, basic residues and deficient in acidic residues.

A heterologous targeting experiment in the apicomplexan *Toxoplasma gondii* demonstrates that the signal peptides of at least one plastid-targeted protein from *B. natans* can substitute for an apicomplexan signal peptide, but the stromal targeting peptide of this protein is not sufficient for import in to the apicomplexan plastid. Whether this is resultis peculiar to this one peptide, or whether the stromal targeting peptides of other plastid-targeted proteins in *B. natans* could substitute for an apicomplexan stromal targeting peptide would be an interesting follow-up experiment. With the eventual completion of the *B. natans* nuclear genome, an abundance of plastid-targeted proteins will be available to expand upon what is currently known about the signal and transit peptides of chlorarachniophytes.
6.5 The chloroplast genome of Bigelowiella natans

Chlorarachniophytes are cells with 4 distinct genomes, a nuclear genome, a mitochondrial genome, a nucleomorph genome and a plastid genome. Currently a large amount of sequence data are available from the nuclear genome in the form of expressed sequence tags and a nuclear genome project of *B. natans* is underway. The nucleomorph genome has been completely sequenced (Gilson et al. 2006), and a mitochondrial genome project of *B. natans* is also in progress (Cavalier-Smith 2006). With the completion of the chloroplast genome and the eventual completion of the nuclear and mitochondrial genomes of *B. natans*, full sequence data will eventually be available from all four genomes of *B. natans*. The chloroplast genome of *B. natans* is the last genome from a distinct group of photosynthetic eukaryotes to be sequenced. A representative chloroplast genomes from haptophytes, heterokonts, cryptomands, apicomplexans, red algae, green algae, charophytes, plants and euglenids have all been sequenced. A complete dinoflagellate chloroplast genome has yet to be sequenced, however the structure of dinoflagellate chloroplast genomes is unusual in being composed of several mini-circle elements, typically with one gene per mini-circle (Zhang, Green, and Cavalier-Smith 1999). The chloroplast genome of *B. natans* is unique in that it is the smallest known chloroplast genome from a photosynthetic organism. Though *B. natans* has fewer protein-coding genes than any other photosynthetic genome, it has only slightly less total coding sequence than the far larger chloroplast genome of *Euglena gracilis*. Hence, the reduced size of this chloroplast genome is likely due to loss of non-coding DNA rather than reduction in coding sequence as I discuss in chapter 5. Furthermore, evidence exists for highly reduced genomes in species of green algae (Manhart et al. 1989), suggesting that perhaps the size of the *B. natans* chloroplast genome may be unusual only in comparison to existing fully sequenced chloroplasts from other photosynthetic organisms. Concatenated phylogenies of plastid genes fail to recover a clade of chlorarachniophytes and euglenids, arguing strongly against the cabozoa hypothesis. It might be argued that phylogenetic analysis of chloroplast genes is not an ideal method for proving the cabozoa hypothesis given that a phylogenetic tree allying the two lineages might simply be the result of a similar endosymbiont in both lineages. Of course, a similar argument could
also be proposed against employing host nuclear data to disprove the cabozoa hypothesis. Since little is currently known about relationships between the five recognized eukaryotic supergroups, it might happen that cercozoa and excavates are in fact sister super-groups, but that the nearest ancestor of both groups does not harbor a green secondary endosymbiont. Although it is agreed that chloroplast data cannot prove the cabozoa hypothesis, it is my contention that it can be used to refute the cabozoa hypothesis for this simple reason: the cabozoa hypothesis rests on the assumption of a common photosynthetic ancestor in both chlorarachniophytes and euglenids. Regardless of the relationship of the plastids of chlorarachniophytes and euglenids to other green algae, the two should at least branch together to the exclusion of chlorophytes. Although an argument could be made for chloroplast loss and replacement events in either or both lineages concealing a single plastid origin in rhizaria and excavates, this hypothesis would be contrary to predictions made by the cabozoa hypothesis regarding independent chloroplast acquisition and targeting events as rare evolutionary events.

I have argued that host nuclear data may not be the best tool for refuting the cabozoa hypothesis, nevertheless a subset of nuclear-encoded genes from chlorarachniophytes and euglenids could be used to independently assess the validity of the cabozoa hypothesis. The majority of both cercozoans and excavates are non-photosynthetic, the only exceptions being phototrophic euglenids among the excavates, and chlorarachniophytes and the enigmatic thecate amoeba *Paulinella chromatophora* within the Rhizaria. A concatenated phylogeny of members of the Rhizaria and Excavata would accordingly depend on genes common to most Rhizaria and Excavata. Hence genes derived from the symbionts of chlorarachniophytes and euglenids would be excluded. I propose that a phylogeny based solely on shared plastid-targeted genes from both chlorarachniophytes and euglenids would provide an independent method of testing the cabozoa. Currently an abundance of published plastid-targeted proteins from both *B. natans* and *E. gracilis* exist. A concatenated analysis similar to the one performed in chapter 5 might be used in a similar fashion to test the cabozoa hypothesis.

Although my analyses suggest that the endosymbiont of *B. natans* is related to a member of the UTC clade rather than a prasinophyte or a deeper branching chlorophyte, the concatenated phylogenies in chapter five suffer from severe undersampling with only
a single representative of each major group of green algae included. In both the full analysis and the reduced one, it is not clear whether \textit{B. natans} branches specifically within the UTC or as sister to this clade. It is my hope that the eventual completion of a larger diversity of green algal and chlorarachniophyte chloroplast genomes will further resolve the exact precise relationships of chlorarachniophytes to this diverse and ancient group of green algae.

6.6 General Conclusions

Although the experiments detailed in this thesis deal with very different aspects of chlorarachniophyte biology, they all relate to the effects of gene transfer on chlorarachniophyte cells. Three Calvin cycle enzymes, class I FBA, FBPase and SBPase show evidence of having been acquired through endosymbiotic gene transfer of the nucleomorph encoded version of this gene to the host nucleus and subsequent re-targeting of their protein products back to the plastid. Other Calvin cycle and pentose phosphate pathway enzymes show evidence of having been acquired through lateral transfer of a eubacterial or eukaryotic gene, as in the case of RPE and TKL. Further evidence is presented for subsequent transfers of the genes encoding these enzymes among eukaryotic taxa. These first two chapters have contributed to a growing body of knowledge on gene transfer and replacement events in eukaryotes with secondary plastids and the prominence of lateral gene transfer in single celled eukaryotes. Understanding the nature of transit peptides of plastid-targeted proteins from \textit{B. natans} also contributes to our understanding of endosymbiotic gene transfer as genes that have been transferred from the symbiont genome to the host often retain their compartment specificity and have to be targeted back to the plastid using targeting pre-sequences. The chloroplast genome is the ultimate source of all endosymbiotic gene transfers involving plastid proteins. The chloroplast genome of \textit{B. natans} is small relative to other photosynthetic eukaryotes, but it is still clearly green algal in the genes that it retains. Phylogenetic analyses of concatenated plastid protein coding genes suggest a relationship between the endosymbiont of chlorarachniophytes and UTC green algae. These studies have contributed to our understanding of chlorarachniophytes and secondary endosymbiosis not only in outlining a possible origin for the endosymbiont of chlorarachniophytes but in
describing several instances of gene transfer from the endosymbiont to the host genome, as well as many instances of lateral gene transfer between the host genome and the genomes of eubacteria and eukaryotes.
6.7 Bibliography


