Transcription of dinoflagellate chloroplast minicircle genes

by

Yunkun Dang

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 (Vancouver) December 2009 © Yunkun Dang, 2009
Abstract

The dinoflagellate chloroplast genome consists of a group of small circular DNA molecules (minicircles), which carry one to three genes. In *Heterocapsa triquetra*, most minicircles carry only one gene except *petD* and *psbE* minicircles, which were found to carry tRNA genes downstream of the protein-coding gene. I used RT-PCR to show that the tRNAs were co-transcribed with the protein-coding genes that preceded them, and cleaved from the precursor before a poly(U) tail was added to the mRNA.

To further understand minicircle transcription and RNA precursor processing, I used RT-PCR, primer extension and Northern analyses to show that some minicircles can produce RNAs larger than themselves. I determined with 5’ RACE the sequence of the processed 5’ ends of several long RNAs, some of which are immediately downstream of the 3’ end of mature mRNAs and tRNAs. I proposed a "rolling circle" model for the minicircle transcription in which transcription would proceed continuously around the minicircle DNA to produce transcripts larger than the minicircle itself. These transcripts would be further processed into discrete mature mRNAs and tRNAs.

I found multiple types of substitutional editing in rRNAs and mRNAs, with A-to-G editing predominating. The editing occurs concomitantly with RNA maturation. I used a bioinformatic approach to generate a secondary structure model of the 16S rRNA. The model suggests that 1) the A-to-G editing mechanism is different from that responsible for animal nuclear A-to-I(G) editing; 2) A-to-G editing increases the conservation of edited sites; 3) the divergent 16S rRNA is a functional component of the chloroplast ribosome.

To prove the presence of a eubacteria-like RNA polymerase (encoded by *rpoA, B, C1 and C2*) in *H. triquetra* chloroplasts, I used degenerate PCR, inhibitor assay, Southern and western analyses to examine the presence of the *rpoB* gene in the chloroplast or nuclear genome. Surprisingly, all the tests gave negative results, suggesting that the *rpoB* might have been lost. On the other hand, by using RT-PCR and 5’ RACE, I cloned an *rpoT* gene which encodes a single subunit RNA polymerase and might be a candidate for the minicircle transcription.
Table of Contents

Abstract ........................................................................................................... ii
Table of Contents ........................................................................................ iii
List of Tables ................................................................................................ viii
List of Figures ............................................................................................... ix
List of Abbreviations ..................................................................................... xi
Acknowledgements ....................................................................................... xii
Co-authorship Statement ............................................................................... xiii

Chapter 1 Introduction .................................................................................... 1
1.1 Introduction to dinoflagellates .................................................................. 2
  1.1.1 General characteristics of dinoflagellates .......................................... 2
  1.1.2 Evolution of dinoflagellates ................................................................. 4
  1.1.3 The chloroplasts of dinoflagellates .................................................... 5
  1.1.4 Minicircles – a unique form of chloroplast genome ......................... 7
1.2 The transcriptional apparatus of plastids ................................................. 12
  1.2.1 Eubacteria-like RNA polymerase ...................................................... 12
  1.2.2 Bacteriophage-like RNA polymerase ................................................. 13
1.3 Post-transcriptional RNA processing and stability .................................. 15
  1.3.1 Overview ............................................................................................ 15
  1.3.2 Chloroplast nuclease .......................................................................... 16
  1.3.3 Chloroplast mRNA maturation ......................................................... 18
Chapter 2 Identification and transcription of transfer RNA genes in dinoflagellate plastid minicircles* ................................................................. 47

2.1. Introduction......................................................................................... 48

2.2 Materials and Methods....................................................................... 49

   2.2.1 Cloning and sequencing of *H. triquetra* genomic minicircles........ 49

   2.2.2 Bioinformatic analysis of non-coding regions.................................. 50

   2.2.3 Reverse transcription-PCR............................................................... 50

2.3. Results.............................................................................................. 51

   2.3.1 *petD, psbE and psbD* unigenic minicircles....................................... 51

   2.3.2 Detection of tRNA genes.................................................................. 52

   2.3.3 Co-transcription of tRNA and protein-coding genes on *psbE* and *petD*
       minicircles............................................................................................ 53

   2.3.4 Conserved and non-conserved features of the tRNAs ..................... 54

2.4. Discussion.......................................................................................... 54

2.5 References........................................................................................ 64
Chapter 3 Long transcripts from dinoflagellate chloroplast minicircles suggest “rolling-circle” transcription* .......................................................... 67

3.1 Introduction ................................................................................................. 68

3.2 Materials and methods ........................................................................ 69

3.2.1 Algal cultures ............................................................................................... 69

3.2.2 RNA extraction and RNA blotting .............................................................. 69

3.2.3 Primer extension .......................................................................................... 70

3.2.4 Reverse-transcription PCR ........................................................................... 70

3.2.5 Mapping of transcriptional initiation sites ................................................... 70

3.3 Results ........................................................................................................... 71

3.3.1 Detection of multicistronic minicircle transcripts ........................................ 71

3.3.2 The long precursor RNAs have processed 5’ ends ................................ .... 73

3.3.3 Single-step cleavage produces the 5’ end of the processed precursor at the same time as the mature 3’ end ................................................................. 75

3.4 Discussion ..................................................................................................... 76

3.4.1 Transcript initiation and termination ........................................................... 76

3.4.2 The rolling-circle transcription model of H. triquetra minicircles ........ ...... 77

3.5 References ................................................................................................... 89

Chapter 4 Substitutional editing of Heterocapsa triquetra chloroplast transcripts and a folding model for its divergent chloroplast 16S rRNA* ................................................................................................................. 93

4.1 Introduction ................................................................................................... 94

4.2 Materials and methods .............................................................................. 95

4.2.1 Algal cultures ............................................................................................... 95

4.2.2 RNA extraction and northern blotting ......................................................... 96

4.2.3 RACE (rapid amplification of cDNA ends) and RT-PCR............................. 96
4.2.4 Comparative analysis of 16S rRNA secondary structure ......................... 97

4.3 Results ................................................................................................................ 97

4.3.1 General properties of RNA editing in *H. triquetra* chloroplasts .............. 97

4.3.2 Substitutional editing of mRNAs ................................................................. 98

4.3.3 16S rRNAs are in pieces .............................................................................. 100

4.3.4 A secondary structure model of the edited 16S rRNA .............................. 101

4.4 Discussion ......................................................................................................... 102

4.5 References ....................................................................................................... 117

Chapter 5 Possible loss of eubacterial-like chloroplast RNA polymerase and cloning of a bacteriophage-like RNA polymerase* ................................ 121

5.1 Introduction ....................................................................................................... 122

5.2 Materials and methods .................................................................................... 123

5.2.1 Northern blotting and Southern blotting .................................................... 123

5.2.2 Primers ......................................................................................................... 123

5.2.3 Inhibitor (rifampicin) test ........................................................................... 124

5.2.4 Cloning of *H. triquetra* rpoT ................................................................. 124

5.2.5 Phylogenetic analysis .................................................................................. 125

5.2.6 Western blotting .......................................................................................... 125

5.3 Results ................................................................................................................ 125

5.3.1 Degenerate PCR could not amplify *rpoB* from *H. triquetra* total DNA.... 125

5.3.2 Southern analysis gave no signal of *rpoB* gene ....................................... 127

5.3.3 Rifampicin did not affect the expression of *psbA* in *H. triquetra* ............ 127

5.3.4 Tobacco RpoB antibody gave no signal for *H. triquetra* total protein extract .............................................................................................................. 128

5.3.5 Cloning dinoflagellate *rpoT* ..................................................................... 128
5.4 Discussion ................................................................................................................. 129

5.5 References ................................................................................................................. 136

Chapter 6 Conclusions and future directions .................................................. 138

6.1 Summary of major discoveries of this thesis ........................................ 139

6.2 The transcription initiation of minicircle and the single subunit RNA polymerase ................................................................. 140

6.3 Crosstalk between RNA editing and RNA processing .......................... 141

6.4 Precise determination of the secondary structure of 16S rRNA ......... 141

6.5 References ................................................................................................................. 143
List of Tables

Table 1.1 The chloroplast genes conserved in plants and algae .................................................29

Table 2.1 Primers used in detection of tRNA transcription ..........................................................57

Table 2.2 Properties of new minicircles from *Heterocapsa triquetra* ..................................57

Table 3.1 The primers used for the outward-directed RT-PCR to detect long minicircle transcripts ..................................................................................................................................80

Table 3.2 The primers used for the RLM-RACE experiments .....................................................81

Table 3.3 Cleavage sites identified with comparative RACE method .........................................82

Table 4.1 Primers used for 5’ RACE of *H. triquetra* chloroplast transcripts ........................106

Table 4.2 Primers used in the 3’ RACE of *H. triquetra* chloroplast transcripts ....................107

Table 4.3 Primers for the RT-PCR of *H. triquetra* chloroplast transcripts to obtain the cDNA sequence data uncovered by the 5’ and 3’ RACE .........................................................108

Table 4.4 Editing of *H. triquetra* chloroplast RNA sequences determined by RACE and RT-PCR ........................................................................................................................................109

Table 4.5 Substitutional editing sites in mRNAs and their influence on translation .............110

Table 5.1 The primers used in semi-quantitative RT-PCR and cloning *rpoB* and *rpoT* .....132
List of Figures

Figure 1.1: The evolution of chloroplasts ................................................................. 26

Figure 1.2: The chloroplast of Heterocapsa triquetra .......................................... 27

Figure 1.3: Examples of the minicircle structure of H. triquetra chloroplast ......... 28

Figure 2.1: Clover leaf structures of tRNA genes ................................................... 58

Figure 2.2: Alignment of Heterocapsa triquetra minicircle sequences upstream of the conserved 9GAG region .......................................................... 59

Figure 2.3: Gene transcription on psbE and petD minicircles .................................. 60

Figure 2.4: Sequence alignment of tRNA-Trp ......................................................... 61

Figure 2.5: Sequence alignment of tRNA-Pro ......................................................... 62

Figure 2.6: Sequence alignment of tRNA-fMet ....................................................... 63

Figure 3.1: Transcription of the psbB minicircle non-coding region tested with RT-PCR. 83

Figure 3.2: Detection of long minicircle transcripts ............................................. 84

Figure 3.3: Determination of the 5’ end of atpA long transcripts ........................ 85

Figure 3.4: Detecting the 3’ end cleavage site of the psbB minicircle transcript ....... 86

Figure 3.5: Detecting the 3’ end cleavage sites for trnW and trnP in petD minicircle transcripts with RLM-RACE ................................................................. 87

Figure 3.6: Scheme describing the transcription mechanism of H. triquetra chloroplast minicircles ................................................................. 88

Figure 4.1: Substitutional editing of the petD minicircle transcripts ..................... 111

Figure 4.2: Editing makes a new stop codon for atpA. The mRNA sequence starts at the transcriptional start site determined by 5’ RACE ............................... 112
Figure 4.3 The 16S rRNA is in pieces. .................................................................113

Figure 4.4 Secondary structure model of post-edited *H. triquetra* chloroplast 16S rRNA .........................................................................................114

Figure 4.5 Comparison of partial secondary structures among *H. triquetra*, *G. theta* and *E. coli* ...........................................................................................................116

Figure 5.1 The evidence supporting the loss of *rpoB* in *H. triquetra* ......................133

Figure 5.2 Phylogenetic analysis of *H. triquetra rpoT* ....................................134

Figure 5.3 The copy number and expression of *H. triquetra rpoT* .......................135
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAR</td>
<td>adenosine deaminase that acts on RNA</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CODEHOP</td>
<td>consensus-degenerate hybrid oligonucleotide primer</td>
</tr>
<tr>
<td>DMSO</td>
<td>dimethyl sulfoxide</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>dsRNA</td>
<td>double stranded RNA</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>HEPES</td>
<td>4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>NEP</td>
<td>nuclear-encoded RNA polymerase</td>
</tr>
<tr>
<td>nt</td>
<td>nucleotide</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>PCP</td>
<td>peridinin chlorophyll a/c binding protein</td>
</tr>
<tr>
<td>PEP</td>
<td>plastid-encoded RNA polymerase</td>
</tr>
<tr>
<td>PNK</td>
<td>T4 polynucleotide kinase</td>
</tr>
<tr>
<td>PPR</td>
<td>pentatricopeptide repeat</td>
</tr>
<tr>
<td>RLM-RACE</td>
<td>RNA ligase-mediated rapid amplification of cDNA ends</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>reverse transcription-polymerase chain reaction</td>
</tr>
<tr>
<td>TAP</td>
<td>tobacco acid pyrophosphatase</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>sodium dodecyl sulfate polyacrylamide gel electrophoresis</td>
</tr>
</tbody>
</table>
Acknowledgements

First I would like to thank my supervisor, Dr. Beverley R. Green for providing me with this opportunity to work on this project. During the past six years, it is her encouragement, guidance and support that enabled me to develop an understanding in this subject and helped me make clear my long term research goal in RNA biology. I would also like to thank my committee members, Dr. Patrick Keeling, Dr. Xin Li, Dr. Tom Beatty and previous member Dr. Tony Griffiths for their brilliant advice on this project and patience to review my thesis.

I am indebted to my colleagues, Dr. Balbir Chaal, Martha Nelson, Dr. Miroslav Obornik, Dr. Meriem Alami, Dr. Songhua Zhu and Linda Lin, who helped me through numerous technical hurdles and inspired me during discussions. I also like to thank Dr. Claudio Slamovits and Dr. Juan Saldarriaga from Keeling Lab for their kind help in bioinformatics. I thank Xuguang Liu and Michael Dong for their assistance in microscopy and sequencing gel electrophoresis, respectively. I thank Qingning Zeng and Jim Guo for generously providing me Arabidopsis materials (RNA and DNA). Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

I wish to thank my sister, Yunxiao Dang and Yunshan Dang, who always kept me away from family responsibilities and encouraged me to concentrate on my study.

I would like to express special thanks to my wife Zilai Zhang for her mental and substantial support to this project and thesis. Without her help and encouragement, this study would have hardly been completed.

Finally and most importantly, I owe my deepest gratitude to my parents, Chenglin Dang and Yuzhen Ren. They bore me, raised me, supported me, taught me, and loved me. To them I dedicate this thesis.


Co-authorship statement

The contribution of the candidate to the publications and/or manuscripts is listed below:


M.J. Nelson and the other authors sequenced minicircles described in the paper. M.J Nelson identified the tRNA genes and wrote the related sections. The candidate corrected the sequence errors in petD and psbE minicircles, performed the tRNA expression experiments and wrote the related section. Dr. B.R. Green conceived the project and helped revise the manuscript.

Chapter 3: Yunkun Dang and Beverley R. Green, Long transcripts from dinoflagellate chloroplast minicircles suggest “rolling-circle” transcription (manuscript has been submitted for publication)

The candidate did all the experiments and wrote the manuscript. Dr. B.R. Green supervised the project and helped revise the manuscript.

Chapter 4: Yunkun Dang and Beverley R. Green, (2009), Substitutional editing of Heterocapsa triquetra chloroplast transcripts and a folding model for its divergent chloroplast 16S rRNA, Gene 442: 73-80

The candidate did all the experiments and wrote the manuscript. Dr. B.R. Green supervised the project and helped revise the manuscript.

Chapter 5: Yunkun Dang and Beverley R. Green, Possible loss of eubacteria-like chloroplast RNA polymerase and cloning of a bacteriophage-like RNA polymerase (manuscript will be submitted for publication)

The candidate did all the experiments and wrote the manuscript. Dr. B.R. Green supervised the project and helped revise the manuscript.
Chapter 1 Introduction
1.1 Introduction to dinoflagellates

1.1.1 General characteristics of dinoflagellates

Dinoflagellates are an important group of phytoplankton widely distributed in marine and fresh waters. About 2000 species of dinoflagellates have been recognized: half of them are photosynthetic while the other half are either parasitic or predatory (Li and Hastings, 1999; Hackett et al., 2004a). Photosynthetic dinoflagellates, often regarded as “algae”, are important primary producers in the ocean, next only to diatoms (van den Hoek et al., 1995). Dinoflagellates often draw our attention by “red tides”, a result of blooming of some species. Red tides make many kinds of marine life suffer from depletion of the oxygen in the water and/or the neurotoxin produced by dinoflagellates. Some dinoflagellates are bioluminescent and can emit blue-green light, though the actual function of this bioluminescence is not very clear (Taylor, 1987). Some dinoflagellates (zooxanthellae) form symbioses with a wide range of protists and invertebrates. The best studied cases are the symbioses between reef-building coral (host) and dinoflagellate Symbiodinium (symbiont): the symbionts provide up to 50% of fixed carbon to the host via excretion, while the hosts are believed to provide a favourable environment and some organic nutrients that cannot be synthesized by the symbionts (Hackett et al., 2004a). The serious environmental problem, coral bleaching, results from changes of marine environments (temperature) that disrupt the mutualistic symbiosis between dinoflagellates and corals.

Dinoflagellates (phylum Dinozoa), together with apicomplexans and ciliates, belong to the infrakingdom Alveolata, which features cortical alveoli or their derivatives beneath the cell membrane (Graham and Wilcox, 2000). Dinoflagellates are characterized by two different flagella: the transverse flagellum girdles the body, often in a groove, and the longitudinal flagellum is oriented perpendicular to the transverse flagellum. Dinoflagellates are encased with a complex covering called the amphiesma, a structure made of continuous outer membranes and a series of closely adjacent, flattened vesicles (alveoli) underneath. Armored dinoflagellates have cellulose or other polysaccharides filling in the cortical vesicles (i.e. thecal plates) that are tightly assembled together to form a strong peripheral skeleton. Thecal plates not only serve
as a way of protection for dinoflagellates, but also an important cue for taxonomy based on their arrangement (i.e. tabulation) (Graham and Wilcox, 2000).

Dinoflagellates are usually haploid during the life cycle and reproduce through fission. However, sexual reproduction exists by fusion of two haploid cells into a planozygote. This zygote either remains motile like a typical dinoflagellate or forms a resting dinocyst (hapnozygote) especially under adverse circumstances, which later produces haploid progeny through meiosis when conditions are favourable (van den Hoek, 1995).

Dinoflagellates contain a large amount of DNA, ranging from 3 up to 200 pg/cell (Rizzo, 2003). In comparison, a human haploid nucleus on average contains only 3.2 pg/cell. Since a eukaryotic algal genome usually harbours about 20,000 genes, the enormous gap between the genome size and the coding regions suggest that most of the dinoflagellate nuclear DNA is structural. Indeed, by randomly sampling a piece of H. triquetra nuclear DNA (230,000bp), McEwan et al. (2008) found that 90% of the sequence contains no recognizable genes and shares no similar feature to any known DNA elements such as transposons or repeats. This might partially explain why the chromosomes are permanently condensed and attached to the nuclear envelope and why this structure remains during mitosis (van den Hoek, 1995). The chromosomes lack nucleosomes and histones, though some basic histone-like proteins (HLPs) do exist (Rizzo, 2003). The HLPs have no phylogenetic link to histones (Wong et al., 2003). The ratio of basic proteins to DNA was estimated to be 1:10, which is in contrast to 1:1 and 1:1.75 ratios in typical eukaryotic nuclei and prokaryotes, respectively (Rizzo et al., 1982). The exact function of HLPs is largely unknown. By observing the distribution of HLPs in different stages of the cell cycle, some early studies suggested that they may regulate gene transcription and maintain chromosome structure (Sala-Rovira et al., 1991). A recent study with in vitro tests of one HLP (HCc3) in Heterocapsa triquetra showed that it functions as dimers and condenses DNA in a concentration-dependent manner (i.e. the higher the ratio of HCc3 to DNA, more condensation of DNAs) (Chan and Wong, 2007). Moreover, in Lingulodinium polyedrum the acetylation of HLPs was also detected and believed to regulate gene
transcription like acetylation of histones in typical eukaryotes (Chudnovsky et al., 2002). Dinoflagellates have eukaryotic G1-S-G2-M cell cycles, but their nuclear envelopes do not disappear during mitosis as observed in plants and animals. Instead, during mitosis the spindle of microtubules develops a channel through the nucleus and directs the segregation of chromosomes (Bhaud et al., 2000).

The organization, regulation and expression of dinoflagellate genes also have some marked characteristics: 1) the promoters of dinoflagellate genes so far analysed do not possess a conventional TATA box or consensus motifs, while other alveolates (ciliates and apicomplexans) do have them (Hackett et al., 2005). 2) Multiple copies of some genes are arranged in tandem repeats. Their transcripts might be polycistronic (e.g. major chlorophyll a/c-containing intrinsic light-harvesting complex proteins, LHCs) or discrete mRNA (peridinin chlorophyll a/c-binding proteins, PCPs) (Le et al., 1997; ten Lohuis and Miller, 1998). 3) Most mRNAs, if not all, are post-transcriptionally capped with a 22 nt spliced leader sequence at the 5’ end (Lee et al., 2007; Zhang et al., 2007). 4) Some mRNAs might be reverse-transcribed into cDNA and integrated into the genome, where they acquire proper conditions (promoter, terminator, etc) to be expressed again. This “recycling” process may happen multiple times for some genes (Slamovits and Keeling, 2008). 5) Some mRNAs may lack poly(A)+ tails (e.g. mRNAs encoding luciferin-binding proteins, LBPs) (Lee et al., 1993). 6) Cytosine methylation is involved in gene expression but the DNA-methylation system might be different from that of green plants based on inhibitor studies (ten Lohuis and Miller, 1998).

1.1.2 Evolution of dinoflagellates

The evolution of dinoflagellates is somewhat symbolized by the evolutionary history of their chloroplasts. At least four important endosymbiotic events were involved in the origin or acquisition of chloroplasts of land plants and algae (Figure 1.1). Primary endosymbiosis probably occurred only once between a non-photosynthetic eukaryote and an engulfed cyanobacterium to form a chimera, which diverged into three lineages (chlorophytes, rhodophytes and glaucophytes). Secondary endosymbiosis occurred between two eukaryotic cells, in which the engulfed photosynthetic eukaryotes were
reduced to three or four envelope-encased plastids (McFadden, 1999; Cavalier-Smith, 2000). Extensive studies on molecular phylogenetics with rRNA and protein sequences have firmly placed dinoflagellates in the clade of alveolates, a late diverging group in eukaryotes (Gajadhar et al., 1991; Fast et al., 2001; Fast et al., 2002; Harper et al., 2005). Recently, alveolates and chromists (containing heterokonts, cryptophytes and haptophytes) have been united and collectively called “chromalveolates” because they all have bipartite leader sequences in their chloroplast-targeted proteins and use chlorophyll c as light-harvesting pigment (Cavalier-Smith, 2003). Although direct phylogenetic evidence that can unambiguously unite all the members of chromalveolates into a monophyletic group is still lacking, many pieces of discrete evidence obtained from chloroplast and nuclear gene data can be put together to suggest that alveolates (including dinoflagellates), heterokonts, haptophytes and cryptophytes might have arisen from a common plastid-bearing ancestor (Baldauf et al., 2000; Yoon et al., 2002; Harper and Keeling, 2003; Harper et al., 2005; Hackett et al., 2007; Patron et al., 2007). Moreover, the replacements of plastid-targeted copies of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and fructose-1,6-Aldolase (FBA) in sampled chromalveolates also support the monophyly of this group (Fast et al., 2001; Patron et al., 2004). In addition, tertiary endosymbiosis, though insignificant for chloroplast evolution, has been found in some dinoflagellates whose chloroplasts were replaced by those of haptophytes, heterokonts, cryptophytes or green algae (Hackett et al., 2004a).

### 1.1.3 The chloroplasts of dinoflagellates

As described above, most photosynthetic dinoflagellate species have secondary plastids, which feature peridinin as a light-harvesting pigment and were derived from a red alga. This type of chloroplast contains chlorophyll a and c2 but lacks chlorophyll b (van den Hoek, 1995). The chloroplast is usually brown since the green chlorophylls are mixed with yellow and red accessory pigments (β-carotene and several xanthophylls, of which the most important is peridinin) (van den Hoek, 1995). Only found in dinoflagellates, peridinin is the major light-harvesting pigment in most photosynthetic dinoflagellates. Peridinin-containing dinoflagellates have two types of
light-harvesting complex: a membrane-intrinsic peridinin-Chl $a/c$ complex (iPCP) and a water-soluble peridinin-Chl $a$ complex (sPCP). Unlike iPCPs, whose protein sequence clearly belongs to the FCP (fucoxanthin chlorophyll $a/c$ binding protein) branch of the light harvesting complex (LHC) superfamily, the sequence of sPCPs shows no similarity to any other pigment protein (Green and Durnford, 1996).

Generally, dinoflagellate peridinin-containing chloroplasts take an irregular tubular shape, which are mainly distributed at peripheral regions of the cell (Wilcox et al., 1982; Horiguchi, 1995). In the case of $H. triquetra$, the chloroplast is made of many flat lobes that together form a spherical network. Most of the lobes are on the surface of this sphere and very few lobes are distributed inside (Figure 1.2). We reasoned that this chloroplast morphology should have an advantage for photosynthesis by increasing the area for capturing light. The number of chloroplasts in a dinoflagellate cell remains unclear. By examining the species *Heterocapsa circulariquasma* with light and electron microscopy, Horiguchi (1995) concluded that only one chloroplast exists in one cell. This is probably true in $H. triquetra$ since those lobes of the chloroplast appear to be connected (Figure 2). It has yet to be determined whether such a shape and distribution of chloroplasts are typical of other peridinin-containing chloroplasts, but the unique chloroplast structure nearly precludes the possibility to purify intact chloroplasts. Indeed, my attempt to purify $H. triquetra$ chloroplasts was largely unsuccessful as most stroma substances were lost in the process.

The conventional peridinin-containing chloroplasts are surrounded by three membranes, none of which is connected to the endoplasmic reticulum (Hackett et al., 2004a). Therefore, dinoflagellates have a distinct mechanism for targeting nucleus-encoded proteins to plastids: plastid-targeting proteins are first inserted in the ER (endoplasmic reticulum), and then transferred to plastids via the Golgi apparatus (Nassoury et al., 2003). Correspondingly, the leader sequences of these plastid-targeted proteins also contain a bipartite structure with an ER-targeting signal peptide at the N-terminus followed by a transit peptide for crossing the chloroplast envelopes. It is worth mentioning that most of the protein transit peptides of dinoflagellates share
a hydrophobic four-residue motif beginning with phenylalanine, which may serve as a specific signal for plastid targeting (Patron et al., 2005).

Although dinoflagellates obtained their second-hand chloroplasts from a red alga, they may lose or replace them with plastids from a heterokont, haptophyte, cryptophyte or green alga (Keeling, 2004). Accordingly, the composition and structure of dinoflagellate chloroplasts would vary based on the source of the plastids.

1.1.4 Minicircles – a unique form of chloroplast genome

Chloroplasts of plants and algae feature multiple identical circular double-stranded DNA molecules endowing their semiautonomy. Their genome size varies from 120 to 220 kb and it encodes 100-250 genes. Although dinoflagellates have retained cytologically recognizable chloroplasts, the plastidial genes of several peridinin-containing dinoflagellates, instead of being located on a large circle as in most plants and algae, are individually located on separate minicircles with the size ranging from 2.2–3.8 kb or 5-10kb in some rare cases (Zhang et al., 1999; Barbrook and Howe, 2000; Zhang et al., 2000; Barbrook et al., 2001b; Hiller, 2001; Nelson and Green, 2005). Accumulating data indicate that minicircles are a distinguishing characteristic of peridinin-containing dinoflagellates. However, not all chloroplast genes are on minicircles. In the dinoflagellate *Lingulodinium polyedrum* the psbA gene is on large DNA molecules that migrated as a band of 50-150 kb in a pulse-field gel (Wang and Morse, 2006b).

A minicircle is comprised of the coding region (ORF) and non-coding regions based on the physical structure (Figure 1.3). The non-coding region of minicircles is usually 0.3-1 kb in length, but it could be as large as 4 kb in some cases (Nelson and Green, 2005). In a given dinoflagellate species, all the minicircles share several conserved cores within the non-coding region. Moreover, in all minicircles of one species the conserved cores and the open reading frames are always oriented in the same direction (see Figure 3, Zhang et al., 2002). Based on these facts, Zhang et al (2002) proposed two models to explain the biogenesis of minicircles: (1) a population of conventional chloroplast DNA carrying multiple genes was subjected to differential gene deletion,
which finally resulted in the generation of small circular DNA molecules carrying different genes. (2) transposition of putative replication elements throughout the chloroplast genome, which resulted in the generation of small circular DNA by homologous recombination. Although it is likely that both models were involved in the biogenesis of minicircles, homologous recombination appears to be more plausible at least for explaining the biogenesis of the aberrant minicircles such as those carrying jumbled genes or empty circles (Barbrook et al., 2001a; Zhang et al., 2001; Nisbet et al., 2004a). Moreover, Nelson and Green (2005) discovered that the large non-coding regions (4 kb) of Adenoides eludens minicircles carry many repeats and some of them can form perfect double hairpin elements (DHEs). Since the DHEs are believed to be mobile elements and have a function in mitochondrial recombination (Paquin et al., 2000), these DHEs found in the non-coding region of minicircles of Adenoides eludens and Symbiodinium suggested that the core regions could function as origins of replication or recombination and gene conversion (Moore et al., 2003; Nelson and Green, 2005).

1.1.4.1 The dinoflagellate chloroplast genome and lateral gene transfer

The first study of dinoflagellate minicircles identified eight protein and two ribosomal RNA genes from H. triquetra (Zhang et al., 1999). Subsequent studies revealed several more genes from several other species. However, to date, only 12 protein, 2 r RNA and 3 tRNA genes altogether have been reported to be located in minicircles (Table 1.1). Since no master circle has ever been identified, these collections of minicircles may represent the entire chloroplast genome, which would therefore be the smallest one in terms of gene number. Most minicircles carry a single gene. However, in Amphilodinium species, two minicircles have been reported that carry two or more genes (Barbrook et al., 2001b; Hiller, 2001; Nisbet et al., 2004b); one carries the petB and atpA genes and the other carries psbD, psbE and psbI. Another case is in Adenoides eludens, where psbA and psbD are together on a large minicircle (10 kb).

The discrepancy between the number of genes in peridinin-containing plastids and other primary or secondary plastids indicates that most of the chloroplast genes have either been lost or relocated to the nuclear genome by endosymbiotic gene transfer, a
process by which the host nuclear genome acquires one or several copies of an organelle gene and redirects the gene products to their original compartment or other destination. This process includes at least two steps: first is the relocation of genetic materials from chloroplasts to the nucleus and next is the acquisition of promoters and targeting signals (reviewed in Bock and Timmis, 2008). For example, most chloroplast genomes of plants and algae encode the α subunit gene (rpoA) of RNA polymerase. However, in a moss *Physcomitrella patens* it was found to be lost in the chloroplast genome and relocated to the nucleus (Sugiura et al., 2003).

Recent studies on the expressed sequence tag (EST) profiles of several dinoflagellate species confirm that many chloroplast genes in dinoflagellate plastids have been transferred to the nucleus, including some genes that are always located on chloroplast genomes in all the other photosynthetic organisms (Table 1.1), ranking them as the “champions of endosymbiotic gene transfer” (Bachvaroff et al., 2004; Hackett et al., 2004b; Tanikawa et al., 2004; Patron et al., 2005). Actually, dinoflagellates not only acquired genes from their organelles but also from other sources. For example, dinoflagellates use an anaerobic proteobacterial form II ribulose 1,5-bisphosphate carboxylase/oxygenase (rubisco) instead of the conventional Form I rubisco for their carbon fixation (Morse et al., 1995). Another example is plastid fructose-1,6-bisphosphate aldolase (FBA) of dinoflagellates and other chromalveolates, which is not the Class I enzyme as in plants and red algae but a Class II enzyme from an unknown eubacterial source (Patron et al., 2004).

### 1.1.4.2 The intracellular location of minicircles

At present many pieces of indirect evidence support the idea that minicircles are located and function in dinoflagellate chloroplasts. (1) All of the genes found on minicircles are conserved in the chloroplast genomes of other plastid-bearing photosynthetic organisms (Table 1). (2) None of the minicircle genes includes a sequence that encodes a recognizable transit peptide or has introns inside, which is different from genes relocated to the nucleus. (3) The codon preference of minicircle genes is different from that of nuclear genes (Codon Usage Database, http://www.kazusa.or.jp/codon/). (4) The EST data so far released do not contain any
counterpart of minicircle genes (Bachvaroff et al., 2004; Hackett et al., 2004b; Tanikawa et al., 2004; Patron et al., 2005), which are crucial for the proper function of photosynthesis. (5) A study on *Symbiodinium* with electron microscopic *in situ* hybridization found that labelled *psbA* antisense RNA probes were located on the thylakoid membranes but not in the cytoplasm (Takishita et al., 2003). Given that no mRNA import into chloroplasts has ever been discovered, the data indicate that the *psbA* genes are located in the chloroplasts. (6) Synthesis of PsbA protein in *Lingulodinium polyedrum* is inhibited by chloramphenicol (an inhibitor of prokaryotic protein synthesis) but not by cycloheximide (an inhibitor of eukaryotic protein synthesis), suggesting that *psbA* is encoded by and its mRNA translated in chloroplasts (Wang et al., 2005).

However, a study on *Ceratium horridum* suggested that the minicircles carrying chloroplast genes are located in the nucleus (Laatsch et al., 2004). The authors extracted a high molecular weight (HMW) DNA from the purified chloroplasts and showed that *psbB* is not on this large DNA but on a minicircle DNA extracted from a nuclear fraction. Since they did not identify any leader sequence directing the minicircle gene products to the chloroplasts and no counterpart is found in the HMW DNA, it remains a mystery how the minicircle genes function in the *C. horridum* chloroplast in this case.

**1.1.4.3 Replication of minicircles**

As most circular DNA, such as most bacterial and chloroplast DNA, uses the θ replication mechanism, it was of course believed that minicircles also use this mechanism. However, a recent study found that the replication intermediates of *H. triquetra* contain large linear DNAs of 6–8 kb, which is in contrast to the 2-3 kb minicircles. Further analyses with 2D-gel electrophoresis and atomic force microscopy revealed that replicating DNA showed characteristics of rolling circle intermediates (Leung and Wong, 2009). Rolling circle replication was first observed in bacteriophage replication and is highly efficient in duplicating template DNA so as to allow the production of multiple copies of products with a single original of replication (Pierce, 2003). However, a study on the copy number of *Amphidinium operculatum*
showed that minicircles are kept at low abundance (about 10 copies/cell) during exponential phase and accumulated (several hundred copies per cell) by early stationary phase (Koumandou and Howe, 2007). This observation contradicts the rationale that rolling circle replication is used to satisfy the demand of cell growth, although it is possible that rolling circle may not be the replication mechanism in all dinoflagellate chloroplasts.

1.1.4.4 Expression of minicircle genes

Generally, gene expression refers to three steps: transcription, post-transcriptional processing and translation. In terms of mechanism of transcription, Nisbet et al. (2008) found that in *Amphidinium* some minicircles have two species of transcripts. The smaller one has a size corresponding to the ORF and appears to be the mature form, whereas the larger one is about the same size as the minicircle, with the 5’ and 3’ end of the transcripts adjacent to each border of the conserved core. Based on these data, the author proposed a conserved motif located at the 3’ end of the conserved core (only one in *Amphidinium carterae* minicircles) as the putative promoter to produce a long primary transcript, which is subsequently processed into a short mature mRNA (Nisbet et al., 2008). However, so far no study has been conducted to determine the 5’ end of the primary transcripts, which would precisely identify the transcriptional initiation sites. In terms of post-transcriptional processing, two unique phenomena, extensive substitutional editing and 3’ end polyuridylation, are worth mention. RNA substitutional editing was first reported in *C. horridum* and then in *L. polyedrum* (Zauner et al., 2004; Wang and Morse, 2006a). In both cases, the editing has a high frequency and encompasses almost all types of transitions and transversions, although A to G transition is the most dominant. However, this is obviously not universal in dinoflagellate minicircle expression as no editing event was discovered in *A. carterae* (Barbrook et al., 2001a). The polyuridylation is also intriguing as such a post-transcriptional addition of poly(U) tail (25-40 nt) has never been reported in chloroplasts. It is obviously not a degradation signal since it is found in the most abundant RNA transcripts (Wang and Morse, 2006a). However, the actual function remains unknown.
Very little is known about the translation of minicircle genes. Wang et al. (2005) reported that the translation of \textit{psbA} in \textit{L. polyedrum} is sensitive to chloramphenicol and regulated by light instead of a circadian clock. However, a subsequent report showed that in \textit{L. polyedrum} the \textit{psbA} is not encoded in a minicircle (Wang and Morse, 2006b). So it remains unclear whether the translation pattern of minicircle genes is same as that of \textit{psbA} in \textit{L. polyedrum}.

1.2 The transcriptional apparatus of plastids

As mentioned above, minicircle genes, though bizarre with respect to their nature, are actively transcribed. Then it comes to the question about the transcriptional apparatus involved in this process. Unfortunately, previous studies gave no hints in the case of dinoflagellates. In order to address this question, reference to some extensively studied plastids from plants and algae is appropriate.

To date two kinds of RNA polymerase active in plastids have been described: PEP (plastid-encoded RNA polymerase) and NEP (nucleus-encoded RNA polymerase) (Smith and Purton, 2002). The PEP genes are found in most plastid genomes so far sequenced, except for those of parasitic plants (Wolfe et al., 1992; Berg et al., 2003; Krause et al., 2003). The NEP, derived from gene duplication of mitochondrial RNA polymerase, is only found in flowering plants.

1.2.1 Eubacteria-like RNA polymerase

PEP is a eubacteria-like RNA polymerase derived from cyanobacterial RNA polymerase, which occurs universally in plastids of all plants and algae. The holoenzyme of eubacteria-like chloroplast RNA polymerase has 5 subunits: \(\alpha\), \(\beta\), \(\beta'\), \(\beta''\) and \(\sigma\), which are encoded by \textit{rpoA}, \textit{rpoB}, \textit{rpoC1}, \textit{rpoC2} and \textit{rpoD}, respectively. Accordingly, the promoters of PEP also have the canonical -10/-35 element similar to that of eubacterial \(\sigma^{70}\)-type promoter (Smith and Purton, 2002). \(\sigma\) factors are involved in recognition of the corresponding promoter. They are released from the holoenzyme when transcription initiation is completed. A variety of genes coding \(\sigma\) factors have been found in plants and algae. Their expression is light-regulated, leaf-specific or
circadian-controlled (Allison, 2000).

In all plastid-containing organisms found to date, including apicomplexans that only contain vestigial plastids (Wilson et al., 1996; Cai et al., 2003), \textit{rpoB, C1 and C2} genes are uniformly located on the plastid genome. Despite several rare cases where \textit{rpoA} is relocated to the nucleus (Sato et al., 2000; Sugiura et al., 2003), it is usually in the plastid genome. In contrast, \textit{rpoD} is always found in the nuclear genome. In most cases, the PEP is crucial for the viability of chloroplasts and plants as it controls the transcription of most, if not all, chloroplast genes. Deletion of PEP genes coding for the core enzyme would result in the absence of chlorophyll and disruption of chloroplast development (Allison et al., 1996; Serino and Maliga, 1998; De Santis-MacIossek et al., 1999). However, in some parasitic plants, the genes encoding the PEP have been lost from the plastid genome and transcription of the genome is therefore mediated by a single-subunit RNA polymerase (Funk et al., 2007).

\textbf{1.2.2 Bacteriophage-like RNA polymerase}

The plastids of higher plants have an additional RNA polymerase (NEP) distinct from the eubacterial PEP. Located in the nucleus, the \textit{NEP} gene encodes a 110 kD single subunit RNA polymerase resembling bacteriophage SP6/T7 RNA polymerase. It is believed to have diverged from mitochondrial RNA polymerase by gene duplication (Lerbs-Mache, 1993; Smith and Purton, 2002). NEP recognizes distinct types of promoters that fall into 3 categories: Class Ia and Ib promoters share a YRT core within a moderately A/T-rich region, though Class Ib has an additional GAA box located about 10-20 nucleotides upstream of the YRT core; Class II does not share any conserved motif with the other two and the -5 - +25 region was proven to be sufficient for specific start of transcription (Weihe and Borner, 1999).

Since the plastid NEP was derived from its counterpart in mitochondria, knowledge from the extensively studied yeast mitochondrial RNA polymerase (Rpo41) may suggest some interesting properties of NEPs. Like NEP, Rpo41 has a highly divergent N-terminus region and a relatively conserved C-terminus region responsible for catalytic function during the process of transcription (Masters et al., 1987). Unlike the
phage RNA polymerase with full function independent of other protein factors, the mitochondrial RNA polymerase of yeasts and mammals requires an additional factor (Mtf1/mtTFB) for the selective recognition of promoters (Kelly and Scarpulla, 2004). To this point, transcription mediated by the single subunit RNA polymerase is similar to that of the canonical eubacteria-like RNA polymerase, both comprising a “core enzyme” and a promoter-recognizing factor (Cliften et al., 1997; Cliften et al., 2000). Moreover, research on the yeast mitochondrial RNA polymerase found that the core enzyme can independently recognize the mitochondrial promoter, while Mtf1 appears to facilitate melting the promoter (Matsunaga and Jaehning, 2004). In higher plants, no protein factor similar to Mtf1 interacting with the NEP has been found in plastids. However, CDF2, a transcriptional factor previously found to interact with PEP, was reported to regulate promoter recognition of NEP in the process of transcribing spinach plastid rDNA (Bligny et al., 2000).

It is believed that the function of plastid NEP in most vascular plants is to transcribe the housekeeping genes, including the rpo genes coding for PEPs, whereas the PEPs mainly (but not exclusively) transcribe the photosynthetic genes (Smith and Purton, 2002). However, NEPs do transcribe photosynthetic genes in some parasitic plants, whose plastid genomes have lost the rpo genes (Berg et al., 2003; Krause et al., 2003). There is no report of the existence of NEP in algal plastids. Indeed, based on the phylogenetic tree, the appearance of NEP is a rather late evolutionary event, having diverged from mitochondrial RNA polymerase after the split of vascular and non-vascular plants (Kabeya et al., 2002; Yin et al., 2009).

Another intriguing aspect of NEP is that of “one gene serving two genomes” (Hedtke et al., 2000). The author found that the RpoT;2 of Arabidopsis has an unique structure at the 5’ end of the ORF: proteins translated from the first initiation codon target to plastids; proteins translated from the second initiation codon (40 codons downstream of the first initiation codon) target to mitochondria. Similar observations were also obtained from Nicotiana sylvestris and moss Physcomitrella patens (Kobayashi et al., 2001; Richter et al., 2002). Moreover, an Arabidopsis mutant in RpoT;2 has been shown to severely impair photosynthesis by decreasing the transcript
level of light-induced genes as well as having a minor effect on the mitochondrial transcription, which indirectly supports the dual targeting of this protein (Baba et al., 2004). However, a study on the PpRpoT1 and PpRpoT2 of *Physcomitrella patens* has refuted the concept of dual targeting. Kabeya and Sato (2005) found that fusion proteins with intact 5’ UTR of PpRpoT are only translated from the second initiation codon, although when translation is forced to start from the first initiation codon, the fusion protein is indeed targeted to plastids. In addition, their reinvestigation of *Arabidopsis* RpoT;2 showed that the dual targeting of this protein is an artifact resulting from the impairment of the native 5’ UTR (Kabeya and Sato, 2005). Therefore, the observations from the RpoT;2 mutant are likely attributable to crosstalk between plastids and mitochondria. Dual targeting, though an incorrect concept according to the available data, still has some implications for the evolution of plastid NEP. Phylogenetic and functional studies support the idea that plastid NEPs gained their function after the angiosperms diverged from gymnosperms (Kabeya et al., 2002). However, the plastid-targeting potential of PpRpoT1 and PpRpoT2 in moss indicates that plastid NEP might have arisen before the gene duplication to form the plastid and mitochondrial lineage. In algae, the mitochondrial RNA polymerase with potential of dual targeting has yet to be tested due to insufficient data about the 5’ end of the transcript.

**1.3 Post-transcriptional RNA processing and stability**

**1.3.1 Overview**

In contrast to bacteria, gene expression in chloroplasts is mainly regulated at the post-transcriptional level. Both bacteria and chloroplasts produce polycistronic RNAs. However, in bacteria, transcription is usually coupled with translation, i.e. translation can be initiated at polycistronic mRNAs. In chloroplasts, however, studies suggest that the primary or immature chloroplast messages are often less active for translation than mature mRNAs (Reinbothe et al., 1993; Shapira et al., 1997; Nickelsen et al., 1999). These studies led to a hypothesis that chloroplast translation is regulated by RNA processing, which is partly true as Bruick and Mayfield (1998) showed that the 5’ end
processing of *C. reinhardtii* *psbA* messages are concomitantly associated with the binding of ribosomes to the 5’ UTR. A recent study further demonstrated that the translation efficiency of mature to immature mRNAs varies in different genes (Yukawa et al., 2007). In general, processed messages show a better translation activity than the unprocessed.

Generally, chloroplast post-transcriptional processing can be divided into several steps: 1) 5’ and 3’ end maturation, 2) intercistronic processing, and 3) RNA editing. If a gene carries an intron, the intron will also be removed during the process of transcript maturation. In this part I will focus on the first two steps and RNA editing will be discussed in detail in the next part.

1.3.2 Chloroplast nucleases

RNA processing is related to endonucleolytic cleavage by endonucleases and trimming by exonucleases. These nucleases are important for chloroplast RNA metabolism as they are involved in the maturation and degradation of RNAs. To date, several endonucleases and exonucleases have been identified:

1.3.2.1 Endonucleases

**RNase E/G**

In *Escherichia coli*, RNases E and G share similar catalytic domains, although the C termini are different. Early studies in bacteria suggested that RNase E has critical function for RNA maturation and degradation and inactivation of this gene would arrest cell division (Ghora and Apirion, 1978; Ray and Apirion, 1981). RNase E/G homologs (encoded by *rne*) have been found on some of chloroplast genomes and have activity similar to that of the *E. coli* enzymes (Schein et al., 2008).

**CSP41a and CSP41b**

CSP41a was first identified as a nonspecific endonuclease that binds to the hairpin structure at the 3’ UTR of spinach *petD* RNAs (Yang et al., 1996). With its substrates, it shows no primary sequence specificity but has strong preference for RNAs carrying
a hairpin structure (Yang and Stern, 1997). As the hairpin structure is crucial for chloroplast RNA stability, CSP41a was believed to function in RNA degradation, which was further confirmed in tobacco (Bollenbach et al., 2003). CSP41b probably is an ortholog of CSP41a. Although the exact function of CSP41b is unknown, a recent study suggested that CSP41a and CSP41b may regulate transcription activity by directing the transcription initiation to a strong promoter region (Bollenbach et al., 2009).

**RNase Z**

RNase Z cleaves the 3’ end of the discriminator (the first unpaired nucleotide at the 3’ end of the tRNA acceptor stem), where the terminal CCA is added. RNase Z plays a critical role for tRNA maturation and has been found in bacteria, archaea, and eukarya. It exists in two forms: a short version of about 250 to 300 amino acids and a long version of about 700 to 900 amino acids (Ceballos and Vioque, 2007). Although RNase Z was first found to target to the nucleus (Schiffer et al., 2002, 2003), in *Arabidopsis* two RNase Z gene products were found to target to chloroplasts (one is mitochondrial and chloroplast dual-targeting) and showed correct cleavage activity (Canino et al., 2009)

1.3.2.2 Exonucleases

Based on their activity, exonucleases can be categorized as 5’-to-3’ and 3’-to-5’. Although the 5’ to 3’ exonuclease activity in *Chlamydomonas* chloroplasts was identified long ago (Drager et al., 1998), no gene has been cloned so far. In contrast, two 3’-to-5’ exonucleases have been identified. One is polynucleotide phosphorylase (PNP, also termed RNase J), which shows both exonuclease and polymerase activity and is responsible for 3’ maturation and degradation (Baginsky et al., 2001; Walter et al., 2002). The other is ribonuclease II/R, which was proven to be responsible for 3’ end maturation of ribosomal RNA (Bollenbach et al., 2005).
1.3.3 Chloroplast mRNA maturation

Unlike nuclear mRNA, chloroplast mRNAs lack a 5’ end cap. The newly synthesized RNA carries a triphosphate group at the 5’ end, which is subsequently removed by 5’ processing. 5’ end processing is important since the 5’ UTR not only controls translation, but also harbour cis-elements affecting the stability of mRNAs (Herrin and Nickelsen, 2004). In general, 5’ end maturation is a combination of actions by cis-elements and trans-factors. For example, C. reinhardii psbD has two types of transcripts: the shorter one with a 47-nt 5’ UTR and longer one with a 74-nt 5’ UTR. With site-directed mutagenesis, Nickelsen et al. detected two cis elements (5’ end and PRB2) required for transcript accumulation (Nickelsen et al., 1999). These cis-elements appear not to form any significant secondary structures. In contrast, a cis-element found in the 5’ UTR of petD was proved to form a stem-loop structure critical for RNA accumulation (Drager et al., 1998), suggesting that the cis elements are gene-specific. Accordingly, the trans factors interacting with these elements to stabilize the RNA are also gene-specific (Drager et al., 1998; Nickelsen et al., 1999).

Bacterial mRNAs usually have a stem-loop structure at the 3’ end, which not only contributes to mRNA stability, but also functions as a rho-independent transcription terminator. Chloroplast mRNAs of plants and algae also possess this structure, but it does not function as transcription terminator, which results in read-through transcription of sequences downstream of the IRs (Stern and Gruissem, 1987; Hayes et al., 1996). However, the 3’ terminus of mature mRNAs ends exactly at the IR, indicating that endo- and/or exo-nuclease activity is required for processing the 3’ overhang of pre-mRNA. One example is the 3’ end maturation of Chlamydomonas atpB mRNA, which was shown to be first cleaved by endonucleases and next trimmed by exonuclease (Stern and Kindle, 1993). However, the endonucleolytic cleavage step might be redundant since mutating or deleting the sequence downstream of the IR did not affect mRNA accumulation and 3’ end formation (Rott et al., 1999).

1.3.4 The chloroplast tRNA maturation

Chloroplast tRNA maturation is largely different from the typical maturation of
bacterial tRNA in that almost no tRNA gene encodes a 3’ CCA sequence. Therefore, a pre-tRNA needs at least three steps of post-transcriptional processing (Mörl and Marchfelder, 2001). First, endonucleases RNase P and RNase Z cleave the pre-tRNA at the 5’ end and 3’ end, respectively, to produce the correct 5’ and 3’ ends. Then CCA is added by tRNA nucleotidyl transferases at the 3’ end to form the mature tRNA. In some cases, introns also need to be removed.

1.3.5 Intercistronic processing

Like bacteria, many chloroplast genes are organized into operons and are co-transcribed into polycistronic pre-mRNAs. In many cases, the polycistronic RNA has to be processed into discrete mature mRNA for active translation. With deletion tests, two cis-elements were identified within the intercistronic spacer between psbT and psbH and between psbH and petB in tobacco, Arabidopsis and spinach (Fei et al., 2007). These elements were predicted to form a stem-loop structure with the cleavage sites at a specific site within the loop region.

1.3.6 Chloroplast RNA stability and polyadenylation

As the purpose of RNA processing is to regulate the moiety of mature RNA that is actively translated, the mRNA must be accumulated and degraded when necessary. As described above, RNA stability is conferred by the interplay of cis elements and trans factors. Among these trans factors, a family of RNA binding proteins termed pentatricopeptide repeat (PPR) proteins have attracted considerable attention since their presence correlates with chloroplast RNA processing and stability in higher plants. For example, in maize chloroplast a PPR protein (PPR10) was shown to bind to the spacers of polycistronic RNAs, defining the boundary of 5’ end and 3’ ends and increasing RNA stability by blocking the progress of exonucleases (Pfalz et al., 2009).

If transcripts of a certain gene are no longer required, degradation will first be initiated by endonucleolytic cleavages at the 5’UTR, 3’ UTR or coding regions. Then the cleaved RNAs are polyadenylated at the 3’ end and degraded by exonucleases (3’-to-5’ and 5’-to-3’). Alternatively, cleaved RNAs can be directly digested by 5’-to-3’ exonucleases, though this pathway appears not prevalent (Bollenbach et al., 2004).
Therefore, as in bacteria, RNA polyadenylation in chloroplasts is not a process to stabilize RNA but a signal to degrade RNA. In bacteria, the enzyme for polyadenylation is PAP 1 (polyA polymerase 1). However, in spinach chloroplasts, such a polymerase was not identified. Instead, polyadenylation was found to be carried out by PNPase, suggesting that the tagging and digestion of target RNAs are undertaken by the same enzyme (Yehudai-Resheff et al., 2001).

1.4 RNA editing

RNA editing is an addendum to the central dogma. Editing increases the genomic plasticity by deleting, inserting or modifying nucleotides so that the RNA sequences are different from their DNA templates. The first evidence of RNA editing featuring uridine insertion/deletion was from mitochondria of trypanosomatid protists over twenty years ago (Benne et al., 1986). Now RNA editing has been found in many different organisms. Based on the way it changes RNA molecules, RNA editing can be categorized as insertion/deletion and substitutional types. However, the mechanism that controls the editing can vary fundamentally.

Substitutional editing has been found in animal nuclei, plant organelles, Physarum mitochondria and dinoflagellate organelles. Substitutional editing does not alter the reading frame of the ORFs, but it still plays a critical role in restoring gene product function, controlling splicing and gene expression. Many different types of substitutional editing have been detected, encompassing nearly all types of base transition and transversion. So far the best studied types are A-to-I and C-to-U editing in animal nuclei, which involve a hydrolytic deamination reaction. However, it is unlikely that all the substitutional editing types proceed via base modification, especially where base transversion is involved.

1.4.1 A-to-I editing in animal nuclei

Adenosine-to-inosine RNA editing, which extensively edits double-stranded viral RNAs to disable the virus, was first found in animal defence against viruses, (Cattaneo et al., 1988). Site-specific A-to-I editing was subsequently identified from RNAs of
glutamate receptors, serotonin receptors and a number of other channel proteins in animal nerve systems (Bass, 2002). The A-to-I editing is catalyzed by adenosine deaminases that act on RNAs (ADARs), which hydrolytically deaminate the C-6 position of the purine ring of adenosine (Polson et al., 1991). As inosine and guanosine share similar molecular geometry, A-to-I editing in RNAs is read as A-to-G editing in vivo, which potentially increases the plasticity of a gene product and provides an effective approach to the gating of channels. In the example of a rat serotonin receptor gene, editing creates as many as 24 protein isoforms and the unedited isoforms showing 20-fold reduced ligand responsiveness compared to the fully edited isoforms (Burns et al., 1997).

Compared with sporadic editing events in coding regions, most editing events have been detected in non-coding RNAs. In human, about 15,000 sites were identified via bioinformatics but nearly all of them are located in Alu repeats (90%) and long interspersed element (LINE) repeats (10%), elements that often reside in introns or mRNA UTRs (Athanasiadis et al., 2004; Blow et al., 2004; Levanon et al., 2004). Direct and indirect evidence supports the conclusion that editing can regulate the expression of genes that harbor these edited introns or UTRs, via alternative splicing, nuclear retention, RNA turnover, and heterochromatic silencing. Moreover, interplays between RNA editing and RNA interference has also been revealed (Nishikura, 2006).

ADAR genes have been found in vertebrates (human, rodent) and invertebrates (worm, fruit fly) but not in plants, fungi, yeast or other known protists (Bass, 2002; Valente and Nishikura, 2005). ADARs contain at least two domains: a double-stranded RNA-binding domain and a deaminase domain, indicating that ADAR needs dsRNAs as substrates. Repetitive non-coding RNAs such as Alu repeats carry most editing events due to their high tendency to form dsRNA structures (Levanon et al., 2004). In the case of mRNA editing, dsRNAs were formed by pairing the edited exons with the edited-site-complementary sequences (ECS) usually located in the downstream introns (Keegan et al., 2001).

As site-specific RNA editing can be directed by ADARs without any additional cofactors, ADAR selectivity is thus dictated by dsRNA secondary structure. In vitro
tests showed that the minimum length of the intermolecular or intramolecular dsRNA region should be no less than 20 bp (Nishikura et al., 1991). Moreover, long dsRNAs (>100 bp) with perfect complementarity are randomly edited, whereas short dsRNAs (20–30 bp) with bulges or mismatches are selectively edited. This indicates that helical structure rather than the primary sequence is important for correct recognition of editing sites (Lehmann and Bass, 1999), although editing preferences for the 5’ bases immediately next to editing sites can be detected in human ADARs (Polson and Bass, 1994; Lehmann and Bass, 2000). Recently, new bioinformatic approaches have been developed to identify a large number of new editing sites from a variety of species, shedding light on the development of a consensus model for specifically edited ADAR substrates in certain species (Sixsmith and Reenan, 2007).

1.4.2 C-to-U editing in mammalian nuclei

Cytidine-to-uridine (C-to-U) editing is another important type of site-specific editing in mammalian nuclei, and is catalyzed by hydrolytic deamination of cytosine at the C-4 position (Johnson et al., 1993). Although it is not as frequent as A-to-I editing in humans, C-to-U editing has been shown to regulate translation of several proteins by changing an amino acid codon to a stop codon (Blanc and Davidson, 2003). The best studied case is apolipoprotein B (apoB) mRNA, where editing converts a CAA codon to UAA stop codon (Chen et al., 1987; Powell et al., 1987). Accordingly, the edited mRNA produces a truncated 48-kD protein (apoB48), compared with a ~100-kD protein (apoB100) encoded by the unedited mRNA. Like most A-to-I mRNA editing events, the apoB C-to-U editing is tissue-specific, which increases the genomic capacity.

Studies of apoB revealed several key requirements for mammalian C-to-U editing. First, the substrates are mature mRNAs (Lau et al., 1991), which is significantly different from A-to-I editing. Second, the editing site selectivity largely depends on an 11-bp primary RNA sequence (mooring sequence) a few bases downstream of the editing site (Driscoll et al., 1993). The cis-elements also include a 3’ spacer and 5’ regulator sequences immediately neighboring the editing site. The mooring, spacer and regulator sequences are believed to fold into a hairpin structure, which positions the
Third, the editing requires an editosome comprising at least two components: cytidine deaminase APOBEC-1 and APOBEC-1 complementation factor (ACF). APOBEC-1 contains a deaminase domain that shares similarity with other nucleoside deaminases (Wedekind et al., 2003). However, overexpression of APOBEC decreased the editing selectivity (Yang et al., 2000), indicating that APOBEC-1 lacks a domain for recognizing the RNA-binding site. In fact, site recognition depends on the ACF, a serine/threonine phosphoprotein that binds both RNA substrate and APOBEC-1 to align the deaminase to editing sites (Mehta et al., 2000). Phosphorylation of ACF can lead to accumulation of ACF in nuclei and increase the editing activity (Lehmann et al., 2006).

### 1.4.3 C-to-U editing in plant organelles

In plant mitochondria or chloroplasts, most RNA editing events are C-to-U changes, though the reverse process, U-to-C changes, can be detected in the plastids of hornworts and ferns (Kugita et al., 2003; Wolf et al., 2004). Early studies centred on the efficiency of RNA editing. Overall, RNA editing is carried out with different efficiency at different sites, i.e. some sites are 100% edited while others are not. Interestingly, in the latter cases, no matter how many different isoforms of RNAs or proteins are produced, only the edited forms of proteins are accumulated (Lu and Hanson, 1994; Phreaner et al., 1996).

As many editing events are involved in changing protein sequences, creating a start codon or stop codon (Shikanai, 2006), RNA editing must be guaranteed to be site-specific. The recognition process should include both cis-element and trans-acting factors. Many in vitro and in vivo strategies have been devised to probe the cis-element for RNA editing recognition (Farre et al., 2007; Hayes and Hanson, 2007; Lutz and Maliga, 2007; Takenaka and Brennicke, 2007). Generally, an ~20-nt sequence upstream and an ~6-nt sequence downstream of the editing site is sufficient to direct site-specific editing, though longer sequences may confer higher efficiency (Reed et al., 2001).
*In vitro* tests also revealed that the *trans*-acting factors that interact with *cis*-elements are proteins instead of antisense RNAs (such as guide RNAs in trypanosomastids) as the editing activity remained when the chloroplast lysate was treated with ribonuclease (Hirose and Sugiura, 2001). With forward genetic approaches, several *trans*-acting factors essential for RNA editing have been identified in *Arabidopsis* (Kotera et al., 2005; Okuda et al., 2006; Okuda et al., 2007; Chateigner-Boutin et al., 2008; Okuda et al., 2009). These proteins belong to the protein family characterized by the presence of multiple copies of a pentatricopeptide repeat (PPR) motif, which is widely distributed in eukaryotes but not in prokaryotes. In the *Arabidopsis* and rice genomes, about 450 and 650 members have been identified, respectively, most of which were predicted to be localized in plastids or mitochondria (Lurin et al., 2004). Other than RNA editing, some of these proteins have been proved to be involved in transcription, splicing, RNA cleavage, translation and RNA stabilization (Schmitz-Linneweber and Small, 2008).

Like its mammalian counterpart, plant organellar C-to-U editing is also a process of deamination rather than nucleotide substitution or transglycosylation (Yu and Schuster, 1995; Hirose and Sugiura, 2001). A zinc-dependent C deaminase activity has been detected in an *in vitro* test of *Arabidopsis* chloroplasts, which is similar to that of APOBEC-1 in mammals (Hegeman et al., 2005). Moreover, analyses of the *Arabidopsis* genome revealed several candidates carrying such a deaminase domain (Faivre-Nitschke et al., 1999). However, so far no experimental evidence has come out to pinpoint the real candidate(s) involved in the editing process.
### 1.5 Objectives

Research on dinoflagellates lags far behind that on other organisms. So far information of the nuclear genome is largely unknown. Although EST profiles provide us with a sketch of gene expression in dinoflagellates, they only represent expression of nucleus-encoded genes, whereas the products of genes with low transcription levels are often absent from these data. No reliable transgenic approach is available in dinoflagellates. Moreover, the branched structure of the dinoflagellate plastid and the hardness of thecal plates prevent the possibility of extracting intact chloroplasts from *Heterocapsa triquetra*. Therefore, to address fundamental questions about minicircle transcription, I turned to molecular biology and bioinformatic approaches.

As described above, previous studies had confirmed: 1) minicircles are indeed the genome of dinoflagellate chloroplasts as they are indeed expressed; 2) the transcripts may contain some modifications such as substitutional editing and polyuridylation. The objective of this project was to understand the transcription and post-transcriptional processing of chloroplast RNAs in dinoflagellate plastids. Specifically, my research goal was to unravel several questions about the dinoflagellate chloroplast gene expression: 1) How is the transcription initiated and terminated? 2) How are the chloroplast transcript precursors processed? 3) What is the possible role of substitutional RNA editing? 4) What is the transcriptional apparatus used for minicircle transcription? Answering these questions will provide a clearer picture of chloroplast gene expression in dinoflagellates.
Figure 1.1: The evolution of chloroplasts

One primary endosymbiotic event forged the common ancestor of red algae, green plants and glaucophytes. Later another three independent endosymbiotic events created a variety of secondary plastid-bearing protists. Some protists (cryptophytes and chororachniophytes) still keep a residual nucleus (nucleomorph) in the secondary chloroplasts.
Figure 1.2: The chloroplast of *Heterocapsa triquetra*
The chloroplast images were obtained by autofluorescence, which was excited with blue light (460-480 nm). Left: stacked images. The size of the scale bar is 10 µm. Right: montage images were taken at a series of planes (1 to 9) from top to bottom. Pictures were taken by Y. Dang.
Figure 1.3 Examples of the minicircle structure of *H. triquetra* chloroplast

The blue arrows show the open reading frame of proteins. The pink arrows show the tRNA genes. The thin black lines show the non-coding regions, which have three conserved core sequences (red boxes). The cores (9G<sub>L</sub>, 9A, 9G<sub>R</sub>) feature nine consecutive guanosines or adenosines in the middle of the conserved sequence (Zhang et al., 1999).
Table 1.1 The chloroplast genes conserved in plants and algae

<table>
<thead>
<tr>
<th>Protein function</th>
<th>Minimal Gene Set (all Photosynthetic plastids)</th>
<th>Additional genes conserved in all red line plastids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photosynthesis-related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP synthase</td>
<td><em>atpA</em>, <em>atpB</em>, <em>atpE</em>, <em>atpF</em>, <em>atpH</em></td>
<td><em>atpD</em>, <em>atpG</em>, <em>atpI</em></td>
</tr>
<tr>
<td>Electron transport between photosystems</td>
<td><em>petB</em>, <em>petG</em></td>
<td><em>petA</em>, <em>petD</em>, <em>petE</em>, <em>petM</em>, <em>petN</em></td>
</tr>
<tr>
<td>Photosystem I</td>
<td><em>psaA</em>, <em>psaB</em>, <em>psaC</em>, <em>psaJ</em></td>
<td><em>psaD</em>, <em>psaE</em>, <em>psaF</em>, <em>psaI</em>, <em>psaL</em></td>
</tr>
<tr>
<td>Photosystem II</td>
<td><em>psbA</em>, <em>psbB</em>, <em>psbC</em>, <em>psbD</em>, <em>psbE</em>, <em>psbF</em>, <em>psbH</em>, <em>psbI</em>, <em>psbJ</em>, <em>psbK</em>, <em>psbL</em>, <em>psbN</em>, <em>psbT</em>, <em>psbZ</em></td>
<td><em>psbV</em>, <em>psbW</em>, <em>psbY</em>, <em>psbX</em></td>
</tr>
<tr>
<td>Transcription/translation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA polymerase</td>
<td><em>rpoB</em>, <em>rpoC1-C2</em></td>
<td><em>rpoA</em></td>
</tr>
<tr>
<td>Ribosomal RNA</td>
<td><em>16S</em>, <em>23S</em></td>
<td></td>
</tr>
<tr>
<td>Large subunit ribosomal proteins</td>
<td><em>rpl2</em>, <em>rpl14</em>, <em>rpl16</em>, <em>rpl20</em>, <em>rpl36</em></td>
<td><em>rpl1</em>, <em>rpl3</em>, <em>rpl4</em>, <em>rpl5</em>, <em>rpl6</em>, <em>rpl11</em>, <em>rpl12</em>, <em>rpl13</em>, <em>rpl18</em>, <em>rpl19</em>, <em>rpl21</em>, <em>rpl22</em>, <em>rpl23</em>, <em>rpl24</em>, <em>rpl27</em>, <em>rpl29</em>, <em>rpl31</em>, <em>rpl32</em>, <em>rpl33</em>, <em>rpl34</em>, <em>rpl35</em></td>
</tr>
<tr>
<td>Small subunit ribosomal proteins</td>
<td><em>rps2</em>, <em>rps3</em>, <em>rps4</em>, <em>rps7</em>, <em>rps8</em>, <em>rps12</em>, <em>rps14</em>, <em>rps18</em>, <em>rps19</em></td>
<td><em>rps5</em>, <em>rps6</em>, <em>rps9</em>, <em>rps10</em>, <em>rps11</em>, <em>rps13</em>, <em>rps16</em>, <em>rps17</em></td>
</tr>
<tr>
<td>Translation factor</td>
<td></td>
<td><em>TufA</em></td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fe homeostasis/FeS Cluster</td>
<td></td>
<td><em>sufB</em> (<em>ycf24</em>), <em>sufC</em> (<em>ycf16</em>)</td>
</tr>
<tr>
<td>Thylakoid translocation</td>
<td></td>
<td><em>secA</em>, <em>secY</em>, <em>tatC</em> (<em>ycf43</em>)</td>
</tr>
<tr>
<td>Protein quality control</td>
<td></td>
<td><em>clpC</em>, <em>dnaK</em>, <em>ftsH</em>, <em>groEL</em></td>
</tr>
<tr>
<td>Mg chelatase</td>
<td></td>
<td><em>ChlI</em></td>
</tr>
<tr>
<td>Cytochrome assembly</td>
<td></td>
<td><em>ccsA</em>, <em>ccsI</em> (<em>ycf44</em>)</td>
</tr>
<tr>
<td>PS I assembly/stability</td>
<td></td>
<td><em>ycf3</em>, <em>ycf4</em></td>
</tr>
<tr>
<td>Acyl carrier protein</td>
<td></td>
<td><em>acpP</em></td>
</tr>
<tr>
<td>Conserved Ycf12</td>
<td></td>
<td><em>ycf12</em></td>
</tr>
</tbody>
</table>

The table was adapted from Green (2004). Redline algae includes rhodophytes as well as chromists and alveolates (whose chloroplasts stemmed from rhodophytes by secondary endosymbiosis). Bold: genes are located in minicircles. Underline: genes are represented in the EST profiles (Bachvaroff et al., 2004; Hackett et al., 2004b; Tanikawa et al., 2004; Patron et al., 2005).
1.6 References


transcription and translation of chloroplast-encoded RNAs in Arabidopsis. Plant Mol. Biol. 69, 541-552.


Hackett, J.D., Yoon, H.S., Li, S., Reyes-Prieto, A., Rummele, S.E., and Bhattacharya, D. (2007). Phylogenomic analysis supports the monophyly of
cryptophytes and haptophytes and the association of Rhizaria with Chromalveolates. Mol. Biol. Evol. 24, 1702-1713.

**Hackett, J.D., Scheetz, T.E., Yoon, H.S., Soares, M.B., Bonaldo, M.F., Casavant, T.L., and Bhattacharya, D.** (2005). Insights into a dinoflagellate genome through expressed sequence tag analysis. BMC Genomics 6, 80-.


two variants of the major basic nuclear protein (HCc) from the histone-less eukaryote Cystheciodinium cohnii (Pyrrhophyta). Chromosoma 100, 510-518.


Chapter 2 Identification and transcription of transfer RNA genes in dinoflagellate plastid minicircles*

* A version of this chapter has been published


** These two authors contributed equally to the study
2.1. Introduction

Dinoflagellates are unicellular eukaryotes and form one of the groups of chlorophyll
a/c-containing algae which acquired plastids by secondary endosymbiosis involving a
red algal endosymbiont (McFadden, 2001; Ishida and Green, 2002). Ecologically, they
are significant primary producers in the world’s oceans, are the source of toxic algal
blooms known as “red tide,” and are associated with reef-building corals as essential
photosynthetic symbionts called zooxanthellae (Taylor, 1987). Dinoflagellates differ
strikingly from all other photosynthetic eukaryotes with respect to their plastid
genomes, which are unique in both gene content and physical organization.

The chloroplast genomes of most photosynthetic eukaryotes have 100-250 genes that
map to a single large circular chromosome of 100-200 kb (Simpson and Stern, 2002),
though at least a fraction of the genomes probably exist as linear concatemers
(Bendich, 2004). However, the plastid genome of dinoflagellates is unique because
most genes found to date are each on a unigenic minicircle of a comparatively tiny size
(2 to 3 kb). Minicircular genes have been found in several species of Heterocapsa
(Zhang et al. 1999, 2001, 2002), Amphidinium operculatum (Barbrook and Howe,
2000; Barbrook et al., 2001), Amphidinium carterae (Hiller, 2001; Zhang et al., 2002),
Ceratium horridum (Zauner et al., 2004), and several strains of Symbiodinium isolated
from various coral host species (Moore et al., 2003). Di- and tri-genic circles of about
the same size as the unigenic circles have been found in several Amphidinium species
(Barbrook et al., 2001; Hiller, 2001), suggesting some selection pressure for small
size. However, the size limit appears to be species-specific, since the dinoflagellate
Adenoides eludens has 5 kb unigenic minicircles and dimeric circles of about 10 kb
(Nelson and Green, 2005).

The dinoflagellate chloroplast genome is also unusual because of the small number of
genes it encodes (Green, 2004; Koumandou et al., 2004). Some of the missing genes
have been transferred to the nucleus: expressed sequence tags (ESTs) have been found
for 18 nucleus-encoded protein-coding genes that are found in the plastid in all other
photosynthetic lineages (Hackett et al., 2004; Bachvaroff et al., 2004). It appears that
the dinoflagellates have taken endosymbiotic gene transfer to the nucleus to a new
extreme. However, there are still a number of conserved protein-coding genes unaccounted for (Green, 2004), and a tRNA gene has only recently been reported (Barbrook et al., 2006a,b). All other plastid genomes sequenced to date carry at least 27 tRNA genes (Wakasugi et al., 1991; Simpson and Stern, 2002; Odintsova and Yurina, 2003). Even non-photosynthetic parasitic plants have at least 17 tRNA genes in their plastid genomes (Lohan and Wolfe, 1998), while the much reduced plastids (apicoplasts) of the apicomplexan parasites such as *Plasmodium* and *Toxoplasma* have 25-33 tRNA genes (Wilson et al., 1996; Denny et al., 1998).

To search for additional chloroplast genes, we screened a large number of clones from a *Heterocapsa triquetra* satellite DNA library. We also investigated the non-coding regions of all publicly available minicircles to look for tRNAs and other small RNA elements, small previously undetected protein-coding sequences, and potential secondary structures such as double hairpin elements (Nelson and Green, 2005). We now report the discovery of tRNA genes on several *H. triquetra* minicircles, and show that they are co-transcribed with other genes on the same circle.

### 2.2 Materials and Methods

#### 2.2.1 Cloning and sequencing of *H. triquetra* genomic minicircles

Using the shotgun library of purified *H. triquetra* satellite DNA (Zhang et al., 1999), 294 additional clones were screened with a probe made to the conserved non-coding region (the 9GAG core found in all minicircles), and the 78 positive clones were rescreened to eliminate the nine genes already found (Zhang et al., 1999). Sequencing of the remaining clones yielded *psbE* and *petD*. The *psbD* sequence was obtained by polymerase chain reaction (PCR) using degenerate primers designed by aligning 21 *psbD* nucleotide sequences and 33 amino acid sequences from various algal and cyanobacterial species. The empty circle (HTcircle6) was assembled directly from shotgun sequences of the satellite DNA library (Zhang et al., 1999). Outward-directed primers and PCR were used to complete the minicircles. The four new minicircles reported in this paper have been deposited in Genbank under accession numbers DQ168850-DQ168853.
2.2.2 Bioinformatic analysis of non-coding regions

The sequences of all completed minicircles, including chimeric and empty circles, were retrieved from the NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/). These included circles from *Amphidinium operculatum*, *A. carterae*, *Heterocapsa triquetra*, *H. niei*, *H. pygmaea*, *H. rotundata*, *Adenoides eludens*, *Protoceratium reticulatum* and several isolates of *Symbiodinium*. The programs blastn and blastx were used to search the non-coding region of every sequence using parameters suitable for picking up small protein-coding genes that might have been missed in earlier analyses. The low-complexity filter was removed for both blastn and blastx searches, and the PAM30 matrix rather than the BLOSUM62 matrix was used in the blastx searches. However, we did not find any new protein-coding genes. The ARAGORN program (https://pcmbioekol-bioinf2.mbioekol.lu.se/ARAGORN1.1/HTML/aragornA.html) was used to search the non-coding regions of all of the dinoflagellate minicircles for tRNAs and tmRNAs (no tmRNAs were detected). ARAGORN uses algorithms to compare input sequences with known tRNA consensus sequences and to check if cloverleaf base pairing is possible (Laslett and Canback 2004). tRNAscan-SE version 1.21 (http://www.genetics.wustl.edu/eddy/tRNAscan-SE/) was also used, picking Mito/Chloroplast as the “source.” tRNAscan-SE identifies candidate tRNAs based on base-pairing and presence of internal promoters, then compares them with covariance models to eliminate false positives; the rate of false positives is estimated to be vanishingly small and the rate of positive predictions to be 99.5% (Lowe and Eddy, 1997). We also tried ERPIN (Easy RNA Profile IdentificatioN) at (http://tagc.univ-mrs.fr/erpin/) but this program did not identify any tRNAs in any of our sequences.

2.2.3 Reverse transcription-PCR

*Heterocapsa triquetra* grown in f/2–Si media (Guillard and Ryther, 1962) at 18° and 50 μmol m⁻² s⁻¹ light intensity was collected near the end of early stationary phase. Cells were broken with a Mini-Beadbeater (Biospec) at 4800rpm for 1 minute with 0.1 mm-diameter beads and total RNA was extracted with RNAqueous –4PCR Kit
(Ambion), then treated with DNase I (Invitrogen) for 30 minutes. For cDNA synthesis, 2 μg of total RNA was transcribed with SuperScript III –RNase H Reverse transcriptase (Invitrogen) using random primers (hexamer, Invitrogen). PCR was done with Taq polymerase (Sigma) and various combinations of primers (Table 2.1). FirstChoice RLM-RACE kit (Ambion) was used for 3’ RACE of petD and psbE, using U tail Adaptor (Table 1) for the first strand cDNA synthesis instead of the 3’ RACE adapter provided in the kit.

2.3. Results

2.3.1 petD, psbE and psbD unigenic minicircles

Two new minicircular genes, petD and psbE, were found by screening an additional 294 clones from the Heterocapsa triquetra satellite DNA library generated by Zhang et al. (1999). The psbD gene was cloned using degenerate primers based on conserved protein sequences. The minicircular sequences were completed using PCR products obtained with outward-directed primers. Because these three genes have already been identified in other dinoflagellate species, their discovery in H. triquetra does not increase the total number of dinoflagellate chloroplast genes reported (Green, 2004). We also found and analyzed an additional minicircle from Heterocapsa triquetra (HTcircle6) which does not encode any identifiable gene; it appears to be an empty circle similar to those found in Amphidinium spp. (Hiller, 2001; Barbrook et al., 2001). Details of these new circles are presented in Table 2.2.

In contrast to the situation in A. operculatum and A. carterae where psbD and psbE are carried on the same 2.4 kb minicircle (Hiller, 2001; Nisbet et al., 2004), in H. triquetra these two genes are each on a unique circle carrying no other protein-coding genes. We could not detect any open reading frame that could have encoded a psbI homolog, which is also found on the psbD/psbE minicircle in A. operculatum. The non-coding regions of the new empty circle and the psbE and psbD minicircles contain normal Heterocapsa triquetra 9G-9A-9G conserved core regions (Zhang et al., 1999; 2002); the petD minicircle has an 8G-9A-8G motif. The gene-coding sequences themselves are closely related to those of other chlorophyll c-containing plastids in the NCBI
database. The alternate start codons of these sequences (Ile, Leu) were also found in several other dinoflagellate chloroplast genes (Zhang et al., 1999; Hiller, 2001; Nelson and Green, 2005).

2.3.2 Detection of tRNA genes

Non-coding sequences of all available dinoflagellate chloroplast minicircles were analyzed with two tRNA detection programs, ARAGORN (Laslett and Canback, 2004) and tRNAscan-SE (Lowe and Eddy, 1997). Both programs detected a putative trnW (tRNA-Trp) gene on several minicircles including the newly sequenced psbE and petD circles, the chimeric circles 3, 4 and 5 of H. triquetra (Zhang et al., 2001), and on both forms of the Heterocapsa pygmaea psbA minicircle (Zhang et al., 1999). Figure 2.1(Left) shows that the H. pygmaea sequence folds into a typical cloverleaf structure, with no obvious mismatches. A putative trnP (tRNA-Pro) gene (Figure 2.1, middle) was also found in these circles by tRNAscan-SE, while a putative trnfM (tRNA-fMet) gene (Figure 2.1, right) was detected by both tRNAscan-SE and ARAGORN only on the H. triquetra petD minicircle. Both trnP and trnfM sequences fold into typical cloverleaf structures (Figure 2.1) although they have one and two mispairings in the stem regions respectively.

Relative gene order is conserved between circles. The H. pygmaea tRNA-Trp genes are found close to the 3’-end of the psbA genes in both of the psbA circles, which differ slightly in the length of the non-coding region (Zhang et al., 1999). In the H. triquetra psbE minicircle, trnW is also found close to the 3’-end of the protein-coding gene (Figure 2.2); in fact, it begins within 12 nucleotides of the end of psbE. In the three H. triquetra chimeric circles, which contain large segments of several protein-coding sequences (Zhang et al., 2001), the trnWs are all located 12nt downstream of psbC fragments. The putative trnP gene is found a short distance downstream of each of the trnW sequences (Figure 2.2). In the H. triquetra petD minicircle, the putative trnfM gene is found just upstream of the trnW gene, and 256 nt downstream of the C-terminus of the petD gene. Alignment of the H. triquetra sequences (Figure 2.2) shows that there is a great deal of sequence relatedness among both coding and chimeric
circles until the beginning of the conserved 9G-9A-9G motif shared by all minicircles of this species (Zhang et al., 2001, 2002).

2.3.3 Co-transcription of tRNA and protein-coding genes on psbE and petD minicircles

To determine whether the tRNA genes are transcribed, and whether they are co-transcribed with the protein-coding genes that precede them, first-strand cDNA synthesis was primed with random primers, and used for PCR with forward primers located in the protein-coding regions of the petD and psbE genes and reverse primers located within the tRNA genes (Table 2.1, Figure 2.3). For the petD minicircle, PCR products of the expected sizes were detected using either petDf1 or petDf2 with tRNAMr, tRNAWr, or tRNAPr (Figure 2.3), as well as using tRNAMf with tRNAPr. The sequences of the PCR products exactly matched the DNA sequences, showing that petD, trnM, trnW and trnP were co-transcribed on a polycistronic mRNA. Similarly for the psbE minicircle, PCR products of the expected size were obtained using psbEf1 or psbEf2 with tRNAWr or tRNAPr, and the sequences matched the genomic sequences. The same results were obtained with cDNA synthesized using tRNAPr or tRNAWr as first-strand primers.

The 3’ ends of some dinoflagellate chloroplast RNA transcripts have recently been shown to be tailed with poly(U) (Wang and Morse, 2006). We found that some processed H. triquetra transcripts also have a poly(U) tail, which provides an effective way to amplify the 3’ end by priming cDNA synthesis with the Utail adapter (Table 2.1). With this cDNA, a PCR product can be produced with primers from within the protein-coding region (petDf1/f2 or psbEf1/f2) and the 3’RACE Outer Primer provided in the kit. Sequencing the PCR products showed that no putative tRNA was included in the poly(U)-tailed transcripts, indicating that the polycistronic RNAs must be subjected to further processing to generate mature tRNAs and tailed mRNAs.
2.3.4 Conserved and non-conserved features of the tRNAs

All the dinoflagellate trnW genes are identical except for the last two nucleotides at the 3’-end of the acceptor stem (Figure 2.4) and have the typical tRNA features of the TψC loop and the anticodon loop (Dirheimer et al., 1995). The D-loop is less conserved with respect to the classic tRNA consensus sequence, but the T to A substitution at position 8 could be compensated by the A to T substitution at position 14. The trnP genes have the consensus features in the anticodon loop, but instead of the highly conserved GTTC sequence in positions 53-56 of the TψC loop they have GTTA (Figure 2.5). It is possible that the A could base-pair with the T substitution in the D-loop. In spite of the fact that the dinoflagellate sequences of both these tRNAs are divergent compared to other plastid tRNAs of the red lineage, they are both predicted to have the normal cloverleaf fold (Figure 2.1).

Although the H. triquetra trnfM sequence folds into a cloverleaf structure (Figure 2.1), it has one mispairing in the anticodon loop and one in the TψC loop, and it does not have the three consecutive GC base-pairs in the anticodon loop found in all other initiator tRNAs (Rajbhandary and Chow, 1995), including those of the other red lineage plastids. While this paper was in review, Barbrook et al. (2006b) published the sequence of a trnfM from the dinoflagellate Amphidinium operculatum. It has three GC pairs in the anticodon loop, although they are not consecutive, and does not conform to the tRNA-fMet consensus at several positions (Figure 2.6). Both dinoflagellate sequences are divergent compared to those of the other plastids. However, our demonstration that the H. triquetra trnfM gene is transcribed suggests that it indeed encodes a functional tRNA and is not a pseudogene.

2.4. Discussion

The chloroplast genomes of red algae, green algae and plants all carry at least 30 tRNA genes. Even the degenerate plastid genomes of the apicomplexans have 25-33 recognizable tRNA genes (Wilson et al., 1996; Denny et al., 1998), while parasitic plants that have lost the ability to photosynthesize still have more than 17 tRNAs (Lohan and Wolfe, 1998). Although the H. triquetra trnW, trnfM and trnP genes are
somewhat divergent compared to their homologs in other chloroplast genomes of the red lineage, they are predicted to fold into the typical clover-leaf conformation and have essential residues (especially well conserved in the anticodon and TψC loops) that suggest that these tRNAs are indeed functional. The tRNA “consensus” appears to be more the exception than the rule in many cases, particularly in organelar genomes (Dirheimer et al., 1995). In fact, these authors comment that there are now no nucleotides that can be considered invariant when mitochondrial tRNAs are taken into account.

All the dinoflagellate tRNA sequences deviate from the standard pattern in the acceptor stem, with the exception of the H. pygmaea trnW and trnP sequences, which are the only ones to have the typical discriminator A at the 3’-end. In fact, within the red lineage plastids, there are a number of cases of mis-pairing in the first 3 nucleotides of the acceptor stem, raising the possibility that there is substitutional editing of these stems, as has been found for some mitochondrial-encoded tRNAs (reviewed in Bullerwell and Gray, 2005). If that is the situation in the dinoflagellate plastids, the editing must occur after the precursors we have detected by RT-PCR are processed to their mature form.

Many protein-coding genes normally found in the chloroplast have been transferred to the nucleus in the dinoflagellates (Bachvaroff et al., 2004; Hackett et al., 2004). Barbrook et al. (2006a) have recently suggested that certain tRNA genes, particularly trnE and trnfM, must be retained by chloroplast genomes for functional reasons, and cannot be replaced by tRNAs imported from the cytosol. If that is the case, then the H. triquetra trnfM should in fact be functional even though it deviates substantially from the standard pattern for initiator tRNA-fMet (RajBhandary and Chow, 1995). The fact that only a few dinoflagellate tRNA genes have been detected suggests that the other tRNA genes may have been transferred to the nucleus, lost, or are simply unrecognizable with current tRNA detection programs. We have found a few sites where there appears to be substitutional editing in the poly(U)-tailed transcripts but not in the uncleaved precursor (Chapter 4.). Substitutional editing of protein-coding and rRNA genes has been found in another dinoflagellate, Ceratium horridum (Zauner et
al., 2004). It is therefore possible that the tRNA sequences are modified after they are cleaved from the precursor.

Acknowledgements

This work was supported by a grant from the Natural Sciences and Engineering Council of Canada (BRG). We thank Dr. M.W. Gray and M.-P. Oudot-Le Secq for helpful discussion, Dr. Oudot-Le Secq for drawing Figure 1, and Drs. M. Turmel and C. Lemieux for sharing their plastid tRNA alignments.
### Table 2.1 Primers used in detection of tRNA transcription

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRNAPr</td>
<td>5’GACCCAAACCAAATACGCTTAAC3’</td>
</tr>
<tr>
<td>TRNAWr</td>
<td>5’GAGGGAACCCTTGTGGACTTGA3’</td>
</tr>
<tr>
<td>TRNAMr</td>
<td>5’CTTTACCTTTTAAACCAAGAGAAGTTTG3’</td>
</tr>
<tr>
<td>TRNAMf</td>
<td>5’CGGTTCCCAAACCTCTCTTGGTT3’</td>
</tr>
<tr>
<td>psbEf1</td>
<td>5’GGCTATAAGTAGAGGTGAACGT3’</td>
</tr>
<tr>
<td>psbEf2</td>
<td>5’TCCATCTCATTACAATCCATCAC3’</td>
</tr>
<tr>
<td>petDf1</td>
<td>5’GCCTTATCTCAAGAGTACCATGTT3’</td>
</tr>
<tr>
<td>petDf2</td>
<td>5’TGAACCTTATAGTGCAGTTGAAC3’</td>
</tr>
<tr>
<td>Utail Adaptor</td>
<td>5’GCGAGCACAGAATTAATACGACTCACTATAGG</td>
</tr>
<tr>
<td></td>
<td>AAAAAAAAAAAAAAAAAABN3’</td>
</tr>
<tr>
<td>3’ RACE Outer Primer</td>
<td>GCGAGCACAGAATTAATACGACT</td>
</tr>
</tbody>
</table>

### Table 2.2 Properties of new minicircles from *Heterocapsa triquetra*

<table>
<thead>
<tr>
<th>Name</th>
<th>Size (nt)</th>
<th>Size of gene (nt, aa)</th>
<th>Start codon</th>
<th>Stop codon</th>
<th>Genbank Acc.No.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>petD</em></td>
<td>2177</td>
<td>477; 158</td>
<td>ATG (Met)</td>
<td>TAA</td>
<td>DQ168853</td>
</tr>
<tr>
<td><em>psbE</em></td>
<td>2195</td>
<td>234; 76</td>
<td>ATA (Ile)</td>
<td>TAA</td>
<td>DQ168852</td>
</tr>
<tr>
<td><em>psbD</em></td>
<td>2629</td>
<td>1067; 355</td>
<td>CTT (Leu)</td>
<td>TAG</td>
<td>DQ168851</td>
</tr>
<tr>
<td>Empty circle HTcircle6</td>
<td>2012</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>DQ168850</td>
</tr>
</tbody>
</table>
tRNA-Trp from *H. pygmaea psbA* minicircle, tRNA-Pro from *H. triquetra* chimeric circle3, tRNA-fMet from *H. triquetra petD* minicircle
Figure 2.2 Alignment of *Heterocapsa triquetra* minicircle sequences upstream of the conserved 9GAG region.

Ht_ch14, 5, 6 are the chimeric circles which contain fragments of several chloroplast genes (Zhang et al., 2001). Arrows show location of primers used for RT-PCR (Table 1), tRNA genes are boxed, conserved nucleotides are starred.
Figure 2.3 Gene transcription on psbE and petD minicircles.

(Upper) Large arrows indicate the coding region and transcript direction of each gene. Small arrows show RT-PCR primer positions (Table 1). Conserved domains of the 9GAG regions are dark gray. Numbers indicate the lengths (bp) of spacers between two adjacent elements. (Lower) RT-PCR products obtained using hexamer-primed cDNA (lanes 1,3,5,7,9, 11,13,14) or total RNA (lanes 2,4,6,8,10,12) templates and the primers indicated below each lane. OP, 3’ RACE Outer Primer from Ambion kit (see Table 1).
### Figure 2.4 Sequence alignment of tRNA-Trp

Alignment of *Heterocapsa* tRNA-Trp sequences with those from plastid genomes of diatoms *Odontella sinensis* and *Thalassiosira pseudonana*; the red algae *Cyandium* (Cy.) caldarium, *Cyanidioschyzon* (Cz.) merolae, *Porphyra purpurea* and *Gracilaria* (Gr.) tenuistipitata; the apicomplexans *Toxoplasma gondii* and *Eimeria tenella* and the cryptophyte *Guillardia* (Gu.) *theta*. Aligned by hand using Bioedit. Consensus tRNA residues are shown on the top line; residues that correspond to this template are highlighted. * indicate residues conserved in all sequences.

<table>
<thead>
<tr>
<th>tRNA consensus</th>
<th>U</th>
<th>Y</th>
<th>A</th>
<th>G</th>
<th>A</th>
<th>R</th>
<th>YUNME</th>
<th>CTTCRA</th>
<th>WC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H. pygmaea pba2</strong></td>
<td>CACCTTAGCAGCTTTTAAR</td>
<td>TAAAATTAAACGCAATTTTTATTAAACCTAGTCGTTCAGTTCC</td>
<td>ACACAGCCTGTTAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H. pygmaea pba2</strong></td>
<td>CACCTTAGCAGCTTTTAAR</td>
<td>TAAAATTAAACGCAATTTTTATTAAACCTAGTCGTTCAGTTCC</td>
<td>ACACAGCCTGTTAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H. triquetra circle1</strong></td>
<td>CACCTTAGCAGCTTTTAAR</td>
<td>TAAAATTAAACGCAATTTTTATTAAACCTAGTCGTTCAGTTCC</td>
<td>ACACAGCCTGTTAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H. triquetra circle4</strong></td>
<td>CACCTTAGCAGCTTTTAAR</td>
<td>TAAAATTAAACGCAATTTTTATTAAACCTAGTCGTTCAGTTCC</td>
<td>ACACAGCCTGTTAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H. triquetra circle5</strong></td>
<td>CACCTTAGCAGCTTTTAAR</td>
<td>TAAAATTAAACGCAATTTTTATTAAACCTAGTCGTTCAGTTCC</td>
<td>ACACAGCCTGTTAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H. triquetra pbe</strong></td>
<td>CACCTTAGCAGCTTTTAAR</td>
<td>TAAAATTAAACGCAATTTTTATTAAACCTAGTCGTTCAGTTCC</td>
<td>ACACAGCCTGTTAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H. triquetra pet2</strong></td>
<td>CACCTTAGCAGCTTTTAAR</td>
<td>TAAAATTAAACGCAATTTTTATTAAACCTAGTCGTTCAGTTCC</td>
<td>ACACAGCCTGTTAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O. sinensis</td>
<td>GGCCTTTAGCTCTGTTT</td>
<td>AGAAGGCTAGCTCCAAATTATGAGTCCAGGTGGTCAACGCTTCGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. pseudonana</td>
<td>GGCCTTTAGCTCTGTTT</td>
<td>AGAAGGCTAGCTCCAAATTATGAGTCCAGGTGGTCAACGCTTCGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cy. caldarium</td>
<td>AGCCTTTAGCTCTGTTT</td>
<td>AGAAGGCTAGCTCCAAATTATGAGTCCAGGTGGTCAACGCTTCGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cz. merolae</td>
<td>AGCCTTTAGCTCTGTTT</td>
<td>AGAAGGCTAGCTCCAAATTATGAGTCCAGGTGGTCAACGCTTCGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. purpurea</td>
<td>AGCCTTTAGCTCTGTTT</td>
<td>AGAAGGCTAGCTCCAAATTATGAGTCCAGGTGGTCAACGCTTCGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gr. tenuistipitata</td>
<td>AGCCTTTAGCTCTGTTT</td>
<td>AGAAGGCTAGCTCCAAATTATGAGTCCAGGTGGTCAACGCTTCGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gu. theta</td>
<td>AGCCTTTAGCTCTGTTT</td>
<td>AGAAGGCTAGCTCCAAATTATGAGTCCAGGTGGTCAACGCTTCGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. gondii</td>
<td>AGCCTTTAGCTCTGTTT</td>
<td>AGAAGGCTAGCTCCAAATTATGAGTCCAGGTGGTCAACGCTTCGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. tenella</td>
<td>AGCCTTTAGCTCTGTTT</td>
<td>AGAAGGCTAGCTCCAAATTATGAGTCCAGGTGGTCAACGCTTCGACG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* indicate residues conserved in all sequences.
Figure 2.5 Sequence alignment of tRNA-Pro

Alignment of *Heterocapsa* tRNA-Pro sequences with those from the plastid genomes in Figure 4, with the addition of two species of *Plasmodium* and the haptophyte *Emiliania huxleyi*. Aligned by hand using Bioedit. Template tRNA residues are shown on the top line; residues that correspond to this template are highlighted. * indicate residues conserved in all sequences.
Alignment of dinoflagellate tRNA-fMet sequences, including the one from empty circle 4 of *Amphidinium operculatum* (Barbrook et al., 2006b), with those from other plastid genomes. Aligned by hand using Bioedit. Template tRNA residues are shown on the top line; * indicate residues conserved in all sequences with the exception of *A. operculatum*; several residues in that sequence (highlighted) either do not conform to the tRNA template sequence or are not shared with *H. triquetra* and all the other species.

<table>
<thead>
<tr>
<th>tRNA consensus</th>
<th>U</th>
<th>YAR</th>
<th>GG</th>
<th>AR</th>
<th>YUN</th>
<th>CRAY</th>
<th>GTCRA</th>
<th>YC</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. triquetra</em></td>
<td>CTTCTGTTGTTTAA-AGGT-AAAAGTTAACATTCCATGATATTAAATT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. operculatum</em></td>
<td>AAATCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. huxleyi</em></td>
<td>GCCTCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>O. sinensis</em></td>
<td>GCCTCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>P. tricornutum</em></td>
<td>GCCTCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. pseudonana</em></td>
<td>GCCTCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>G. theta</em></td>
<td>GCCTCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. merolae</em></td>
<td>GCCTCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>P. purpurea</em></td>
<td>GCCTCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>T. gondii</em></td>
<td>GCCTCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. tenella</em></td>
<td>GCCTCAAGGCTCTAAAAGTTAGTT-AGTCTCATAGCCGAAGGT-ATAGGGCCACTTCTTTTATGAGATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.6 Sequence alignment of tRNA-fMet

Alignment of dinoflagellate tRNA-fMet sequences, including the one from empty circle 4 of *Amphidinium operculatum* (Barbrook et al., 2006b), with those from other plastid genomes. Aligned by hand using Bioedit. Template tRNA residues are shown on the top line; * indicate residues conserved in all sequences with the exception of *A. operculatum*; several residues in that sequence (highlighted) either do not conform to the tRNA template sequence or are not shared with *H. triquetra* and all the other species.
2.5 References


Chapter 3 Long transcripts from dinoflagellate chloroplast minicircles suggest “rolling-circle” transcription*

*A version of this chapter has been submitted for publication

Yunkun Dang and Beverley R. Green, Long transcripts from dinoflagellate chloroplast minicircles suggest “rolling-circle” transcription. (manuscript in preparation)
3.1 Introduction

Dinoflagellates are a diverse group of protists. About one half of dinoflagellate species are photosynthetic and serve as important primary marine producers, though they are best known as culprits of red tides. Most photosynthetic dinoflagellates have chloroplasts derived from a red alga via secondary endosymbiosis (Takishita et al., 2004; Takishita et al., 2005), which feature three envelope membranes and a unique light harvesting pigment, peridinin. The chloroplast genomes are also unique in that the genes, instead of being located on a large circular DNA of 100-200 kb, are individually harbored in small circular DNAs of about 2-3 kb (Zhang et al., 1999; Barbrook and Howe, 2000). A few minicircles are larger (up to 10 kb) and carry 2-4 genes, but genes found together are not in the same clusters as in other chloroplast genomes or cyanobacterial genomes (Barbrook et al., 2001; Hiller, 2001; Nelson and Green, 2005; Nelson et al., 2007a). So far, only 17 genes have been identified on minicircles. They encode 12 proteins, 2 rRNAs and 3 tRNAs. Apart from these genes, several hypothetical proteins have also been identified from *Amphidinium carterae* and *Ceratium horridum* (Laatsch et al., 2004; Barbrook et al., 2006). However, compared with the size of other chloroplast genomes, the dinoflagellate chloroplast genome is undoubtedly the smallest (Green, 2004). A number of missing chloroplast genes have been relocated to the nuclear genome, partially explaining this severe reduction of the chloroplast genome size (Bachvaroff et al., 2004; Hackett et al., 2004).

The transcription of minicircle genes has been tested in a variety of dinoflagellate species (Zhang et al., 1999; Barbrook et al., 2001; Takishita et al., 2003; Zauner et al., 2004; Barbrook et al., 2006; Wang and Morse, 2006). For any protein-coding gene, the most abundant type of transcripts features a poly (U) tail at the 3’ end and is believed to be the mature mRNA (Wang and Morse, 2006; Nelson et al., 2007a). In addition, some low abundance species of long RNAs are occasionally detected. In *A. carterae*, several genes have large transcripts with sizes comparable to the minicircles (Nisbet et al., 2008). RT-PCR results showed that the 5’ end and 3’ end of these transcripts border each end of a conserved core sequence in the non-coding region. Nisbet et al.
(2008) proposed that a conserved motif GAAACGACA in the core may serve as the promoter, from which a long precursor RNA is transcribed and subsequently processed to a short mature one. In *Heterocapsa triquetra*, polycistronic transcripts of the *petD* and *psbE* minicircles can be detected, where the protein-coding gene is co-transcribed with the downstream tRNA genes (Nelson et al., 2007). In the present study, we used northern analyses, primer extension and RLM-RACE to show that some minicircle transcripts of *H. triquetra* are larger than their templates (minicircles), suggesting a unique “rolling-circle” transcription mechanism. Moreover, by using a RLM-RACE method, we provide evidence to show how the long transcripts are processed into mature mRNAs.

### 3.2 Materials and methods

#### 3.2.1 Algal cultures

*Heterocapsa triquetra* (CCMP 449) was obtained from the Provasoli–Guillard Culture Center for Marine Phytoplankton (Boothbay Harbor, ME) and grown in f/2–Si media at 18° on a 12-h light/12-h dark cycle at 50 µmol m⁻² s⁻¹ light intensity. Unless specified, cells were usually collected at 14 days after inoculation (at approximately 0.5-1×10⁶ cell/mL), corresponding to the early stationary phase.

#### 3.2.2 RNA extraction and RNA blotting

For RNA blotting, 14-day cultures were used for total RNA extraction following the instructions of RNAqueous –4PCR Kit (Ambion) with a few modifications. Cells resuspended in the Lysis/Binding Solution were broken with a Mini-Bead beater (Biospec) at 4800 rpm for 30 sec with 0.1 mm-diameter beads. Eluted RNAs were treated with TURBO DNA-free™ (Ambion) for 30-60 minutes to digest residual DNAs in the RNA samples. RNA electrophoresis, transfer and hybridization followed Sambrook et al (1989). Each lane contains 25µg total RNA. Radioisotope-labelled primers were prepared with Strip-easy Labelling Kit (Ambion).
3.2.3 Primer extension

Primer extension was carried out following Sambrook et al (1989). T4 polynucleotide kinase (PNK) and SuperScript III reverse transcriptase were purchased from Ambion and Invitrogen; 30 µg total RNA was used for each reaction. The primer extension products were separated on a 4% polyacrylamide gel containing 7 M urea, run at 30 mA for 16 hr so that high molecular-weight bands could be separated.

3.2.4 Reverse-transcription PCR

About 2 µg of total RNA was used for cDNA synthesis with Super Script III –RNase H Reverse transcriptase (Invitrogen) and random primers (hexamer, Invitrogen). Primer sequences and primer combinations are listed in Table 3.1. PCR products were directly sequenced so that the results represent the population of cDNAs.

3.2.5 Mapping of transcriptional initiation sites

The method takes advantage of a modified procedure for RNA ligase-mediated rapid amplification of cDNA ends (RLM-RACE) (Bensing et al., 1996). In this study, we used the FirstChoice RLM-RACE kit (Ambion) and followed the manufacturer’s instructions with minor modifications. First, no calf intestine phosphatase treatment was used for the total RNA. To prepare the (T+) cDNA sample, about 10 µg total RNA was treated with tobacco acid pyrophosphatase (TAP) to remove the 5' triphosphate group. One half of the reaction product (about 5 µg treated total RNA) was then used in an RNA ligation reaction that enzymatically added a synthesized RNA linker (provided in the kit) to the 5’ end of the RNAs. A fraction of the linker-ligated RNAs (2µg) was made into cDNAs with random primers. For the (T-) cDNA sample, 5 µg total RNA, without any TAP treatment, was directly used in the RNA ligation reaction and the subsequent cDNA synthesis. For each minicircle transcript, two rounds of PCR were performed. The first round PCR used either (T+) or (T-) cDNAs as template, with the outer anchor primer 5OP and an outer gene-specific primer (inside or close to the coding region). Each PCR product was diluted 200-1000 times to serve as the template for a second round of PCR. PCR products from (T+) and
(T-) samples were separated side by side on agarose gels. The bands were purified and sequenced to determine the precise 5’ end position. Reverse primers used for the 5’ RACE of each gene are listed in Table 3.2.

3.3 Results

3.3.1 Detection of multicistronic minicircle transcripts

In our previous study, the 5’ UTRs of mature chloroplast mRNAs of *H. triquetra* were determined to be 40 nt on average (Dang and Green, 2009). However, the 5’ ends of these mRNAs are unlikely to represent the transcription initiation sites because the simplified RACE method used in that study can only detect processed 5’ ends carrying a monophosphate group (See methods and Dang and Green, 2009). Moreover, analysis of the genomic sequences immediately upstream of the 5’ ends of the mature mRNAs (putative promoters) showed no consensus elements such as -35 and -10 motifs (data not shown). To determine the 5’ UTR of the *psbB* precursors, we used RT-PCR with the reverse primer inside the coding region and the forward primers in various locations (Figure 3.1). Surprisingly, each primer pair gave a product of the estimated size, even the primer pair psbBr and psbBf. In other words, the 5’ UTR of *psbB* transcripts can extend across the full non-coding region and well into the 3’ end of the coding region. Although the DNA templates (minicircles) are closed circles, the transcripts cannot be circular. Therefore, the RNA detected by psbBr and psbBf primers could represent a long multicistronic RNA containing at least two open reading frames (ORFs), provided that the transcription is not started and terminated inside the ORF. We expanded our test to seven other protein-coding minicircles (*psbA, psbC, psbD, psaA, psaB, petD* and *atpA*) and found that all of them can produce some transcripts whose 5’ UTR can extend into the 3’ end of the preceding ORF. Therefore, it is possible that a minicircle can produce some species of RNA carrying more than one repeat of the DNA template.

To further confirm that minicircles can produce long precursor RNAs, we picked four minicircles (*psbB, psbC, psaA* and *psaB*) for primer extension experiments and chose gene-specific reverse primers located about 300 bp downstream of the start codons.
(Figure 3.1, Figure 3.2A). Each lane contains multiple bands, suggesting that a minicircle could produce more than one species of RNA with UTRs larger than those of the mature mRNAs (40nt on average). By referring to the sequence of the minicircle templates (Figure 3.2A, inserted table), we reasoned that the 5’ UTR of these RNA species can be divided into two categories (Figure 3.2A). In one type, the UTR only covers the non-coding regions. In some cases, a very small portion of the ORF upstream of the non-coding region may also be included. This type can be found in all tested samples. The other type is a transcript carrying an additional intact ORF upstream of the non-coding region. This type was only found in psbB and psaA lanes. In the psbB lane, four bands of about 1 kb, 1.1 kb, 1.3 kb and 3 kb can be detected.

Since the primer is 341 nt downstream of the start codon, these RNA species should carry 5’ UTRs of about 0.66 kb, 0.76 kb, 0.96 kb and 2.6 kb, respectively. As the non-coding region of psbB is 840 bp, the 5’ end of the two larger RNAs could span the entire non-coding region. As for the largest band (3kb), its 5’ UTR (2.6kb) should contain one full non-coding region (0.8kb), one full ORF (1.4kb) and a preceding partial non-coding region (0.4kb).

Since the long transcripts are in low abundance, it is possible that the labelled primers could exceed the available binding sites and hybridize with other RNAs carrying similar sequences, leading to false signals (Bensing et al., 1996). To eliminate this possibility, we performed RNA blotting for these four minicircles to confirm that minicircles can produce transcripts longer than themselves. To assess the difference in expression, we used total RNA purified from cells in the light or dark phase for northern blotting. As shown in Figure 3.2B, the pattern of hybridization is similar in both cases, except that the overall expression is slightly stronger in the light phase than the dark phase.

Under low exposure conditions, only one strong band could be seen in each sample lane (data not shown). These bands represented the mature RNAs since the sizes are similar to the corresponding ORFs. By extending film exposure time, some weak bands of higher molecular weight could be detected in psbB, psbC and psaA lanes (Figure 3.2B, arrowheads). The psbB lane has one weak band (~2.7kb) and 2 faint
bands (~4kb and 4.5kb) above the main band (~1.5kb). The psbC lane has two faint bands (~3.2kb and 4.2kb) above the main band (~1.5kb). These weak bands in psbB and psbC lanes are longer than the size of the corresponding minicircles (2.3kb and 2.2kb, respectively). The psaA lane has a weak band of about 3kb, larger than the main band (~2.3kb) and comparable to the size of the minicircle. Based on the evidence from RT-PCR, primer extension and northern blotting, it is clear that minicircles of H. triquetra chloroplasts can produce some transcripts that are longer than the minicircles themselves. No long transcripts were detected for psaB, although the RT-PCR and primer extension results show that there are transcripts that span the non-coding region.

3.3.2 The long precursor RNAs have processed 5’ ends

To characterize the 5’ ends of the long RNAs, we used a modified RNA ligase-mediated (RLM)-RACE method which is able to distinguish between primary and processed transcripts (Bensing et al., 1996). In chloroplasts or mitochondria, newly initiated transcripts carry a triphosphate group at the 5’ end, while processed transcripts only carry a monophosphate group (e.g. Swiatecka-Hagenbruch et al., 2007). Since the T4 RNA ligase can only catalyze the ligation reaction between a 3’ hydroxyl group and a 5’ monophosphate group, a synthetic oligoribonucleotide can only be ligated with processed RNAs. Therefore, the RLM-RACE can only provide the 5’ end sequence of processed RNAs (Figure 3.3A). To distinguish between the primary and processed 5’ end, tobacco acid pyrophosphatase (TAP) is introduced to produce a monophosphate group suitable for the RNA ligation. As the 5’ RACE products from TAP-treated (T+) RNAs represent both the nascent and processed products while those from TAP-untreated (T-) RNAs only represent processed ones, the additional band(s) from T(+) lanes therefore should show the transcription initiation sites.

Figure 3.3B shows the example of the atpA minicircle. The first round PCR results showed that both (T+) and (T-) cDNAs gave a similar result (Figure 3.3B, left gel). Sequencing these PCR products showed that the 5’ end was exactly the same as that of the mature RNA tailed with poly(U), suggesting that the 5’ end of the mature mRNA
is derived from RNA processing rather than from the transcription initiation site. We tested the other 9 protein-coding genes and they gave the same results as atpA. Therefore, we conclude that the 5’ ends of mature chloroplast RNAs are derived from cleavage of larger precursor RNAs.

In an attempt to detect the 5’ end of the atpA precursor RNAs (present in trace amounts), we performed second round PCR with the first-round PCR products (diluted 200 to 500 times) as the templates, and reverse primers at various locations (Figure 3.3A). The right gel image of Figure 3.3B shows the results of second round PCRs. In Lanes 1 and 2, the sequence was that of the mature atpA 5’ end since atpAnr1 is actually inside the 5’ UTR of the mature mRNA. Theoretically, low abundance long RNAs could be singled out from the mature RNAs if the reverse primers are located upstream of the mature atpA RNAs (atpAnr2 and atpAnr3). Lanes 3 and 4 (using primer Anr2) gave several products but all were found to be a false amplification except the largest one (outlined). Compared to the mature atpA RNA, the sequence of this RNA has a long 5’ UTR (943bp), which extends through most of the non-coding region of the atpA minicircle and ends only 143 bp from the 3’ end of the atpA ORF (Figure 3A). This long 5’ UTR must have been processed from a longer RNA as it can be amplified from both T+ and T- samples. Lanes 5 and 6 also only have one correct product (outlined), which has the same 5’ end sequence as found in Lane 3 and 4. Therefore, the atpA minicircle has at least two forms of transcript, a short mature one and a long processed precursor. The (T+) and (T-) samples always gave the same pattern, suggesting that very few primary (unprocessed) atpA transcripts exist.

Similar results have been obtained for petB, petD, psbB, psbE, psaA and psaB minicircles. Each of these minicircles could produce at least one species of precursor with a long 5’ UTR (Table 1). As all the results can be obtained from both (T+) and (T-) cDNA, these data represent the sites where precursor RNAs are cleaved from primary transcripts initiated at unknown sites.
3.3.3 Single-step cleavage produces the 5’ end of the processed precursor at the same time as the mature 3’ end

The RLM-RACE data in Table 3.3 suggest that the 5’ ends of the *psaA* and *psbB* long precursors are only a few bases from the 3’ ends of the ORFs. For example, 5’ RACE of the *psbB* minicircle transcripts gave just one band, which represents the processed 5’ end as it shows in both the T(+) and T(-) lane (Figure 3.4A). The 3’ end of the mature mRNA was determined by taking advantage of the 3’ poly(U) tail, which is a hallmark of mature chloroplast mRNAs (Wang and Morse, 2006, Nelson et al., 2007). Alignment of sequence data from both the 5’ RACE and 3’ RACE against the DNA sequence clearly showed that the 3’ end of the mature mRNA is next to the 5’ end of the precursor RNA (Figure 3.4B). Combined with the evidence that *psaA* and *psbB* minicircles can produce some precursor RNAs longer than the DNA template, it therefore appears that the mature 3’ end and the precursor 5’ end are generated by a single endonucleolytic cleavage of a polycistronic RNA. The precursor must then be cleaved again to give the mature 5’ end, with a putative uridylyltransferase adding the poly(U) tail to the 3’ end.

A previous study showed that *petD* and *psbE* minicircles also carry several tRNA genes that are co-transcribed with the upstream protein-coding gene (Nelson et al, 2007). As with all the other genes investigated here (e.g. Figure 3.1), outward-directed RT-PCR showed that transcripts of these minicircles could span the entire non-coding region including the tRNA genes and extend into both ends of the protein-coding gene (data not shown). RLM-RACE analysis of *petD* minicircle transcripts gave two species of precursor RNA (Fig 3.5A). Since the bands appeared in both the T(+) and T(-) lanes, the 5’ ends of these two precursor RNAs were also generated from RNA cleavage rather than transcription initiation. Alignment of the 5’ RACE sequencing data with the *petD* minicircle sequence showed that one of the processed long RNA 5’ ends is next to the predicted 3’ end of *trnW* and the other to that of *trnP* (Figure 3.5B). Therefore, the first step of tRNA maturation should be an endonucleolytic cleavage that produces the mature 3’ end of the tRNA. In the case of *trnP* this leaves an extended 5’ end that terminates at the 3’ end of the previous gene (*trnW*). In the case
of the psbE minicircle, two cleavage sites were identified (Table 3.3). One is at 1757 bp upstream of the psbE start codon and right at the 3’ end of the predicted trnP gene. The other is 1548 bp upstream and inside the conserved non-coding region but not near any recognizable ORF.

3.4 Discussion

3.4.1 Transcript initiation and termination

Long chloroplast transcripts, almost the same size as the minicircles, have previously been reported in A. carterae (Nisbet et al. 2008). However, no transcripts covering the entire non-coding region of the minicircle were detected. In this study, we show that in another dinoflagellate, H. triquetra, minicircle transcripts can extend across the non-coding region (Figure 3.1) and can be even longer than the DNA template (Figure 3.2). The sizes of the high molecular weight bands in the psbB and psbC lanes on RNA blots imply that the transcripts could carry two complete ORFs separated by a non-coding region (Figure 3.2B). The psaA band of about 3 kb is the same size as the minicircle, but primer extension shows a faint band of about 4 kb which would imply the existence of a transcript with a second partial ORF.

Chloroplast mRNAs in plants usually require 5’ processing as the 5’ UTRs of the primary transcripts are often larger than those of the mature mRNAs. This means that two 5’ ends can be detected: one corresponding to the transcription initiation site and the other to the processed (mature) end (Westhoff and Herrmann, 1988; Swiatecka-Hagenbruch et al., 2007). In this study, we were unable to find the transcription initiation site(s) for any tested minicircle. However, using RLM-RACE we were able to detect long precursor molecules that had clearly been processed from even longer transcripts because they had 5’ monophosphate rather than 5’ triphosphate ends (Figure 3.3). The 5’-end of the mature mRNA was generated by a second endonucleolytic cleavage (Figure 3.2). In some cases (psaA, psbB minicircle transcripts), the 3’ end of the mature mRNA exactly meets the 5’ end of the processed precursor RNA (Fig 4), suggesting that a single nucleolytic cleavage creates the
mature 3’end and the 5’end of precursors. This implies that the initial transcript (so far undetected) was longer than the minicircle DNA.

In plants and algae, almost all the chloroplast mRNAs have a small invert repeat (hairpin) in the 3’ UTR. In *E. coli*, this stem-loop structure can function as a rho-independent transcription terminator. However, in plant and algal chloroplasts it mainly functions in mRNA stability and is very inefficient for transcription termination, which results in 3’ end read-through of chloroplast RNAs (Stern and Gruissem, 1987; Rott et al., 1996). Transcription termination in chloroplasts is not well understood (Monde et al., 2000), but may involve protein factors as shown in mammalian mitochondria (reviewed in Scarpulla, 2008). In the case of *H. triquetra* it is unlikely that stem-loop structures are required, since most minicircles do not contain any recognizable inverted repeats (Zhang et al., 2002).

### 3.4.2 The rolling-circle transcription model of *H. triquetra* minicircles

Based on the data presented in this study, we put forward a model to describe the transcription mechanism of *H. triquetra* minicircles (Figure 3.6). Since the templates are closed DNA (minicircles), we hypothesize that the transcription of a minicircle, once initiated, proceeds along the circular DNA for many rounds to produce polycistronic precursor RNAs. These polycistronic precursors are then cleaved to give processed pre-mRNA precursors that have a long 5’ UTR but the mature 3’ end sequence. The formation of the mature mRNA involves further 5’ cleavage to produce a molecule with a short 5’ UTR (about 40 nt) and the addition of the 3’ poly (U) tail.

We term this unique mechanism “rolling-circle” transcription since it is reminiscent of the replication of bacteriophages and dinoflagellate minicircles (Leung and Wong, 2009). In this model, due to the physical characteristic of minicircle, the high accumulation of minicircle gene transcripts could be realized with a weak promoter because the minicircle can theoretically produce numerous repeats of themselves with a single initiation. In fact, transcription could be initiated at random sites, rather than downstream of a strong promoter.
In prokaryotes, pre-tRNA usually needs several ribonucleases (RNase E and several exonucleases) to trim the protruded 3’ end to expose the CCA arm (Mörl and Marchfelder, 2001). In chloroplasts and mitochondria, since most tRNA genes do not encode CCA at the 3’ end, the processing at the 3’ end is a single step of endonucleolytic cleavage (Mörl and Marchfelder, 2001). The endonuclease for processing the 3’ end of chloroplast tRNAs has been identified as a ribonuclease Z (Schiffer et al., 2002; Ceballos-Chavez and Vioque, 2005; Canino et al., 2009). Upon cleavage, it leaves a hydroxyl group at the 3’ end of the tRNA and a phosphate group at the 5’ end of the remaining transcript (Ceballos and Vioque, 2007). In *H. triquetra*, the *trnW, trnM* and *trnP* genes in *psbE* and *petD* minicircles do not encode a 3’ CCA. Moreover, our data show that the endonucleolytic cleavages at the 3’ ends of *trnW* and *trnP* leave a monophosphate group at the 5’ end of the remaining transcript, since the 5’ end phosphate group is the prerequisite for the ligation of the RNA linker. Taken together, these observations suggest that the processing of tRNA 3’ ends in dinoflagellate chloroplasts is also likely to be carried out by an RNase Z.

In plant chloroplasts, chloroplast mRNA maturation requires removal of the read-through sequence after the 3’ end stem-loop structure with the aid of an exonuclease (Stern and Gruissem, 1987; Hayes et al., 1996). In *H. triquetra*, the 3’ ends of *psbB* and *psaA* mRNAs are solely generated by endonucleolytic cleavage. Since the biochemistry of the cleavage is similar to that of tRNAs (products carry a 5’ phosphate group), we hypothesize that the 3’ end of mRNAs could be generated by the same endonuclease. Since a deletion mutant of chloroplast tRNase Z in Arabidopsis causes a lethal effect, Canino et al. (2009) suggested that tRNAase Z might control the 3’ end maturation of mRNAs as well as tRNAs. If this is the case in dinoflagellate chloroplasts, RNAase Z could also be the endonuclease responsible for mRNA processing (Figure 3.6).

In summary, we used several different methods to prove that minicircles could produce large mRNA precursors and identified the cleavage sites in some precursors. These data suggest that the transcription of dinoflagellate minicircles follows a unique rolling-circle manner. Moreover, the processing of the 3’ end of mRNAs is very
similar to those of tRNAs, suggesting that the single step of endonucleolytic cleavage for mRNAs and tRNAs might be carried out by the same ribonuclease.
Table 3.1 The primers used for the outward-directed RT-PCR to detect long minicircle transcripts.

For each minicircle, the forward and reverse primers were used. The psbE minicircle transcripts were not applied for this test since the psbE code region is too small.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>psbBnf1</td>
<td>ACACATGCAATTTGCTTTGGATT</td>
</tr>
<tr>
<td>psbBnf2</td>
<td>TTGGGGTTTTGAGGAAAATCTCT</td>
</tr>
<tr>
<td>atpAf</td>
<td>GCT GCT TAC AGT GGT GCT GCT</td>
</tr>
<tr>
<td>atpAr</td>
<td>GGA ACC ATT GCG TCA ACA GAT</td>
</tr>
<tr>
<td>petBf</td>
<td>CGT GGT GGC TTC AGT GTT GG</td>
</tr>
<tr>
<td>petBr</td>
<td>CGA AAC CTG AAG CAC CTT GGA</td>
</tr>
<tr>
<td>petDf</td>
<td>TCA GAA TCC ATT CAG AAG ACC AAT CA</td>
</tr>
<tr>
<td>petDr</td>
<td>AGG CCA TGC AGG TTC ACC AT</td>
</tr>
<tr>
<td>psaAf</td>
<td>TGC TGG TAT TCC TAG CGC AAA</td>
</tr>
<tr>
<td>psaAr</td>
<td>ACC GAG ATC CTG GTT GAG AAT</td>
</tr>
<tr>
<td>psaBf</td>
<td>TTC GCT TGT GAT GGT CCA GGT A</td>
</tr>
<tr>
<td>psaBr</td>
<td>CCA CCT GAA GTA GCT GAA ACG</td>
</tr>
<tr>
<td>psbAf</td>
<td>TCG TGT TCT CAG CTC CAC A</td>
</tr>
<tr>
<td>psbAr</td>
<td>GTG GAG CTG AGA AAG CAA CGA</td>
</tr>
<tr>
<td>PsbBf</td>
<td>ATT CCG TTC TGG TCC ACA GCT C</td>
</tr>
<tr>
<td>psbBr</td>
<td>CAA TGC CAG AAA GCT GCT AAG</td>
</tr>
<tr>
<td>psbCf</td>
<td>CTG GTC TGT CCA AGG TGG TTG</td>
</tr>
<tr>
<td>psbCr</td>
<td>AAA CAC CGC AGA CGA AAT AAG A</td>
</tr>
<tr>
<td>psbDf</td>
<td>TTG GAT GGC TGT CCA AGA TCA</td>
</tr>
<tr>
<td>psbDr</td>
<td>TGA ACC AAC CAC CAG CAG CTA A</td>
</tr>
<tr>
<td>psbEr</td>
<td>TTG GTT GAA GTT TGG AAC ACC</td>
</tr>
</tbody>
</table>
Table 3.2 The primers used for the RLM-RACE experiments.

The cDNA was made from T(+) or T(-) RNAs with random primers (see methods). The first round PCR was performed with outer primers (e.g. psbBr and 5OP) and second round with inner primers (e.g. psbBnr2 and 5IP). The 5IP and 5OP were provided by the RLM-RACE kit (Ambion).

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>atpAnr3</td>
<td>TTC AAG GCA AAT TGC ATG TGT T</td>
</tr>
<tr>
<td>atpAnr2</td>
<td>AAC CCA AAT CTG AGC TCC ACA A</td>
</tr>
<tr>
<td>atpAnr1</td>
<td>TCC ATA TAC GAA GAT TTT CTT TTC G</td>
</tr>
<tr>
<td>petBnr3</td>
<td>CGT TTG ACA AGA AAC CGC AAA</td>
</tr>
<tr>
<td>petBnr2</td>
<td>TTT TTA AAG GAC TTG CGG CTT G</td>
</tr>
<tr>
<td>petBnr1</td>
<td>TTC AAA CAC CCA ATC ACC ATC A</td>
</tr>
<tr>
<td>petDnr3</td>
<td>CCA AGGCAAATTGCA TGT GTT</td>
</tr>
<tr>
<td>petDnr2</td>
<td>CCC AAA TCC CAC AAA AAC ACA</td>
</tr>
<tr>
<td>petDnr1</td>
<td>TCC GAT ACC TTC GGC ACT CTT</td>
</tr>
<tr>
<td>psaAnr2</td>
<td>CCC CCA ATT TCT GAT CCA CA</td>
</tr>
<tr>
<td>psaAnr1</td>
<td>GAA CCG ACG GGA AAC AAC AA</td>
</tr>
<tr>
<td>psaBnr3</td>
<td>TTT TTA AAG GAC TTG CGG CTT G</td>
</tr>
<tr>
<td>psaBnr2</td>
<td>CAA AAC CCC CAA TTT CTG ATC C</td>
</tr>
<tr>
<td>psaBnr1</td>
<td>TAC CAG AAT TTT CCC CCT CTT C</td>
</tr>
<tr>
<td>psbBnr2</td>
<td>GGC AAA TTG CAT GTG TTT TGG A</td>
</tr>
<tr>
<td>psbBnr1</td>
<td>AGA TTT CCC CAA AAC CCC AAA T</td>
</tr>
<tr>
<td>psbEnr2</td>
<td>TAC CAT AAG CTG CGC AAA GGA</td>
</tr>
<tr>
<td>psbEnr1</td>
<td>AGC ACC ACC ACT GCC CTT AAT</td>
</tr>
<tr>
<td>psbEnr3</td>
<td>CCC CCA ATT TCT GAT CCA CA</td>
</tr>
<tr>
<td>5OP</td>
<td>GCT GAT GGC GAT GAA TGA ACA CTG</td>
</tr>
<tr>
<td>5IP</td>
<td>CGC GGA TCC GAA CAC TGC GTT TGC TGG CTT TGA TG</td>
</tr>
</tbody>
</table>
Table 3.3 Cleavage sites identified with comparative RACE method

<table>
<thead>
<tr>
<th></th>
<th><code>psaA</code></th>
<th><code>psaB</code></th>
<th><code>psbB</code></th>
<th><code>psbE</code></th>
<th><code>petB</code></th>
<th><code>petD</code></th>
<th><code>atpA</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>5'-UTR length (bp)</td>
<td>803</td>
<td>790</td>
<td>838</td>
<td>1757</td>
<td>1548</td>
<td>1091</td>
<td>1294</td>
</tr>
<tr>
<td>determined by RACE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-coding region length (bp)</td>
<td>806</td>
<td>888</td>
<td>840</td>
<td>1961</td>
<td>1545</td>
<td>1712</td>
<td>1086</td>
</tr>
</tbody>
</table>
Figure 3.1 Transcription of the psbB minicircle non-coding region tested with RT-PCR

Left: Structure of psbB minicircle with the approximate positions of primers (arrow heads) and open-reading frame (heavy arrow). Right: RT-PCR results with different combinations of primers. The cDNA was made with total RNA and random primers. Total RNA was used as control to test the residual DNA in the RNA sample. All PCR products were purified and sequenced.
Figure 3.2 Detection of long minicircle transcripts.

A: Primer extension analyses of *psbB*, *psbC*, *psaA* and *psaB*. For each sample, 30µg total RNA was used for cDNA synthesis with labelled gene specific primers; reaction products were resolved on a 3.5% denaturing polyacrylamide gel containing 7M urea. Interpretation of the results is shown beside the gel image. Arrows: full and partial open-reading frames (ORFs), thin lines: non-coding region (NC). Table: lengths of minicircle, coding and non-coding regions (bp), names and positions of primers (first nucleotide of the start codon is +1).

<table>
<thead>
<tr>
<th></th>
<th><em>psbB</em></th>
<th><em>psbC</em></th>
<th><em>psaA</em></th>
<th><em>psaB</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Minicircle</td>
<td>2286</td>
<td>2310</td>
<td>3005</td>
<td>3121</td>
</tr>
<tr>
<td>Coding region</td>
<td>1446</td>
<td>1383</td>
<td>2199</td>
<td>2331</td>
</tr>
<tr>
<td>Non-coding region</td>
<td>840</td>
<td>947</td>
<td>906</td>
<td>700</td>
</tr>
<tr>
<td>Primer</td>
<td><em>psbBr</em></td>
<td><em>psbCr</em></td>
<td><em>psaAr</em></td>
<td><em>psaBr</em></td>
</tr>
<tr>
<td>Position of primer</td>
<td>+341</td>
<td>+348</td>
<td>+354</td>
<td>+371</td>
</tr>
</tbody>
</table>

B: RNA analysis of *psbB*, *psaA*, *psbC* and *psaB* minicircle transcripts. Total RNA was purified from cells growing in mid-light phase (L) or mid-dark phase (D). Total RNA (25 µg) was separated on a 1% formaldehyde agarose gel with Riboruler (Fermentas) as size marker. The labelled probes for each gene (about 0.5kb) were complementary to part of the protein-coding region.
Figure 3.3 Determination of the 5’ end of atpA long transcripts

A: Schematic of atpA minicircle and experimental strategy. Two sets of cDNA were made with the RLM-RACE method, using either TAP-treated (T+) or TAP-untreated (T-) total RNAs. PCRs were carried out with various combinations of gene-specific reverse primers and forward primers complementary to the RNA linker (5’ RACE Outer Adaptor Primer, 5OP; 5’ RACE Inner Adaptor Primer, 5IP). Since the 5’ end of newly initiated RNAs carries triphosphates, it can only be detected in (T+) samples.

B: First and second round PCR. To detect the low abundant pre-RNAs, nesting PCRs were carried out with conditions shown below each lane. The first round PCR products (1st PCR) were diluted 200-500 times with distilled water and served as the templates for a second round of PCR (2nd PCR) with 5IP and nesting reverse primers (atpAnr1, 2 or 3). All products were sequenced; correct products are outlined. The length of the long precursor determined by RLM-RACE is indicated by a thin arc in A.
Figure 3.4 Detecting the 3' end cleavage site of the psbB minicircle transcript

The 5’ end of the long psbB precursor was determined by RLM-RACE as in Figure 3.3. A: Schematic of psbB minicircle with positions of primers used for 5’ and 3’ RACE (arrow heads). B: Second round PCR with psbBnr2 and 5IP primers. The first round PCR used psbBr or psBnr1 primer. Both T+ and T- gave a single product. Black and red arc outside the minicircle map in A show the approximate span of 5’ RACE and 3’ RACE (psbBf) relative to the DNA template. C: Alignment showing the sequences of the long precursor RNA detected by RT-PCR (with psbBf and psbBr) compared to 3’ RACE and 5’ RACE results. The sequences of the poly(U) tail and the RNA linker (enzymatically added in 5’ RACE) are indicated by blue- and green-filled boxes. The red box indicates the stop codon of the psbB ORF.
Figure 3.5 Detecting the 3’ end cleavage sites for trnW and trnP in petD minicircle transcripts with RLM-RACE.

A: Schematic of petD minicircle and the approximate position of primers (arrow heads). The 5’ RACE was carried out as described in Figure 3.3. First round PCR used petDr or petDnr1 primers; second round PCR used petDnr2 primer. Second round PCR results are shown in the middle (gel image). Both bands were sequenced and the partial chromatograms are shown on the right. The approximate spans of the 5’ UTRs for these two species of precursor RNA are shown in the red and black arcs (upper and lower gel bands respectively).

B: Partial alignment of 5’ RACE sequence with petD minicircle long transcripts detected by outward-directed PCR as in Fig 3.1. Black boxes, predicted tRNA genes (Nelson et al., 2007); green bars, RNA linkers.
Figure 3.6 Scheme describing the transcription mechanism of *H. triquetra* chloroplast minicircles

Possibly initiated randomly, the transcription continuously goes along the circular DNA template to produce long primary transcripts. The transcripts are subsequently cleaved by some endonuclease (possibly RNase Z) to generate pre-mRNA and/or pre-tRNA, both of which require further trimming at the 5’ end and other processing such as 3’ polyuridylation for mRNAs and 3’ CCA addition for tRNAs.
3.5 References


Chapter 4 Substitutional editing of Heterocapsa triquetra chloroplast transcripts and a folding model for its divergent chloroplast 16S rRNA*

*A version of this chapter has been published

Yunkun Dang and Beverley Green, (2009), Substitutional editing of Heterocapsa triquetra chloroplast transcripts and a folding model for its divergent chloroplast 16S rRNA, Gene 442: 73-80
4.1 Introduction
Photosynthetic dinoflagellates are major primary producers in the ocean, second only to diatoms (Field et al., 1998). The typical chloroplasts of peridinin-containing dinoflagellates are surrounded by three membranes and possess a unique genome: genes are individually located on separate minicircles with the size usually ranging from 2-10 kb (Zhang et al., 1999; Barbrook and Howe, 2000; Barbrook et al., 2001; Hiller, 2001; Nisbet et al., 2004; Nelson and Green, 2005). Most minicircles carry just one gene, although a few carry 2 or 3 genes. Compared with a conventional chloroplast genome that carries 100-200 genes, so far only 17 genes and a few unidentified open reading frames in total have been found in a variety of dinoflagellate species (Howe et al., 2008). The EST profiles suggest that this severe genome reduction is a result of massive endosymbiotic gene transfer to the nucleus (Bachvaroff et al., 2004; Hackett et al., 2004).

Site-specific substitutional editing is a post-transcriptional process of base conversion, which differentiates the RNA molecules from their DNA templates and therefore increases the genomic plasticity. In animal nucleus-encoded RNAs, the most common type of substitutional editing is from adenosine(A) to inosine(I) (Bass, 2002), although cytidine(C)-to-uridine(U) editing is occasionally found (Smith, 2007). Usually detected as A-to-G editing in cDNA due to the similarity of geometry between guanosine and inosine, A-to-I editing has only been proven experimentally in metazoans (Sixsmith and Reenan, 2007). Although A-to-I editing was first detected in the coding regions of mRNAs (Sommer et al., 1991), it is much more frequent in non-coding regions, especially in the transcripts containing Alu repeats in humans (Levanon et al., 2004). In plant chloroplasts and mitochondria, C-to-U editing is the most common type of editing, but U-to-C changes can occasionally be detected. Usually occurring in coding regions, many C-to-U or U-to-C editing sites are crucial for protein function by creating start or stop codons, or changing amino acid sequences (Shikanai, 2006).

Dinoflagellate mitochondrial and chloroplast mRNAs also undergo substitutional editing. In contrast to plant organelles and animal nuclei, they contain many types of
base conversion, some of which occur very frequently, particularly A-to-G editing. The chloroplast minicircle transcripts (psaA, psbB, psbE and 16S rRNA) of *Ceratium horridum* collectively contain several hundred editing sites, which encompass seven editing types (Zauner et al., 2004). A similar level of editing was found for the chloroplast transcripts in another dinoflagellate, *Lingulodinium polyedrum* (Wang and Morse, 2006), but another dinoflagellate, *Amphidinium operculatum*, did not show any evidence of editing (Barbrook et al., 2001). Substitutional editing is even more common in dinoflagellate mitochondrial transcripts (Lin et al., 2002; Zhang and Lin, 2005; Jackson et al., 2007; Zhang et al., 2008; Zhang and Lin, 2008). Editing in both types of dinoflagellate organelles is similar: many types of base transition and transversion exist, with A-to-G editing the most frequent. Although most editing events were found in mRNA, in *Karlodinium micrum* editing of fragmented mitochondrial LSU rRNAs was also detected (Jackson et al., 2007).

*H. triquetra* is the dinoflagellate species where chloroplast minicircle genes were first discovered (Zhang et al., 1999). To date, only 15 chloroplast genes have been found in this species: 10 protein-coding genes, two rRNA genes and three tRNA genes (Zhang et al., 1999; Nelson et al., 2007). Moreover, some minicircles carry jumbled chloroplast gene fragments, which appear to be the result of recombination between different minicircles (Zhang et al., 2001). Here we report an in-depth examination of substitutional editing in this species, which showed that it occurs at a late stage in RNA maturation, increases sequence conservation, and suggests how a very divergent 16S rRNA sequence could still fold properly and function effectively in plastid translation.

### 4.2 Materials and methods

#### 4.2.1 Algal cultures

An axenic culture of *Heterocapsa triquetra* (CCMP 449) was obtained from the Provasoli–Guillard Culture Center for Marine Phytoplankton (Boothbay Harbor, ME) and grown in f/2 –Si media at 18° with 12-h light/12-h dark cycles at 50 μmol m⁻² s⁻¹
light intensity. Cells were usually collected at 28 days after inoculation (cell density approximately $10^6$ cell/mL), corresponding to the early stationary phase of growth.

### 4.2.2 RNA extraction and northern blotting

Total RNA was extracted from the 28-day cultures collected at the midpoint of the light phase or the dark phase with methods described by Nelson et al. (2007). About 20µg total RNA of each sample was used for northern analysis. RNA electrophoresis, transfer, hybridization and washing followed Sambrook et al. (1989). Specifically, a pre-hybridization buffer (50% formamide, 5×SSPE, 2×Denhardt’s reagent, 0.1% SDS, 100µg/mL sheared salmon sperm DNA) was used in the pre- and hybridization at 42˚. Then the blots were washed 3 times at room temperature for 15 min in 1×SSPE and 0.5% SDS. If background was strong, the blots were further washed in 0.2×SSPE and 0.1% SDS at 50˚ for 20-40 min. Radioisotope-labelled probes were prepared with Strip-easy Labelling Kit (Ambion) following the instructions.

### 4.2.3 RACE (rapid amplification of cDNA ends) and RT-PCR

The dinoflagellate chloroplast transcript features a poly(U) tail of 20-30nt at its 3’ end, which is likely a hallmark of RNA maturation (Wang and Morse, 2006; Nelson et al., 2007). To obtain the 5’ and 3’ end sequences of mature mRNAs, we simplified the RNA ligase-mediated (RLM) RACE method described in the FirstChoice RLM-RACE kit (Ambion). First, 10µg total RNA from the light-phase sample was ligated with a synthetic RNA linker at the 5’ end. Second, cDNA was synthesized from 4µg of treated total RNA primed with Utail Adaptor (Table 4.2). Since the Utail Adaptor carries a string of adenosines complementary to the poly(U) tail, the synthesized cDNAs theoretically represent the mature forms of chloroplast RNAs and can be used for both 5’ and 3’ RACE. For 5’ RACE of the 23S and 16S rRNA, we used a similar approach except that random primers (hexamers) replaced the Utail Adaptor for the cDNA synthesis.

Except for *petB, petD* and *psbE*, the 5’ and 3’ RACE results did not cover the entire mRNA. The missing sequence was obtained by RT-PCR with the same cDNAs as templates. All PCR products were directly sequenced so that the results represent the
entire population of transcripts. The primers for 5’ RACE, 3’ RACE and RT-PCR are listed in Table 3.1, 3.2 and 3.3, respectively.

### 4.2.4 Comparative analysis of 16S rRNA secondary structure

The post-edited sequence of *H. triquetra* 16S rRNA was aligned with that of *Escherichia coli*. Conserved regions were plotted based on the secondary structure model of *Escherichia coli* 16S rRNA (comparative RNA web site, http://www.rna.ccbb.utexas.edu/) (Cannone et al., 2002). The structures of short intervening helical segments were predicted using the co-variation approach, where G:C, A:U and G:U base-pairing is conserved even though the primary sequences are not conserved (Cannone et al., 2002). Secondary structures of divergent regions were drawn manually with reference to the results of the MFOLD prediction program (Zuker, 2003) and the corresponding region of the *E. coli* structure model.

### 4.3 Results

#### 4.3.1 General properties of RNA editing in *H. triquetra* chloroplasts

Using RT-PCR and RACE techniques, we determined the full sequences of all 10 mature chloroplast mRNAs (*psbA, psbB, psbC, psbD, psbE, psaA, psaB, petB, petD* and *atpA*) and the 16S rRNA. All the mature mRNAs carry poly(U) tails at their 3’ ends. Moreover, 16S rRNA also has a poly(U) tail, which agrees with the results from *L. polyedrum* and *A. carterae* (Wang and Morse, 2006). 23S rRNA does not seem to have a poly(U) tail as 3’ RACE gave no products. Although we were unable to obtain the full 23S rRNA sequence, we did succeed in amplifying its 5’ end sequence (675nt). All the sequences above have been deposited in GenBank (FJ491245-56).

Except for *psbA* and *psaB*, all genes have at least one editing site. There are 7 types of substitutional editing (Table 4.4). Compared with *C. horridum* and *L. polyedrum* (Zauner et al., 2004; Wang and Morse, 2006), *H. triquetra* has one new type of editing, an A-to-U change, but it occurs at a very low frequency. As in *C. horridum* and *L. polyedrum*, A-to-G changes are the most abundant, accounting for 87% and 52% of editing sites in rRNAs and mRNAs, respectively. The U-to-C editing ranks
second but only occurs in mRNAs. Unlike the other two dinoflagellates where the editing of chloroplast rRNAs and mRNAs has roughly the same frequency, in *H. triquetra* the editing of chloroplast rRNAs is 9.6 times more frequent than that of mRNAs (Table 4.4).

### 4.3.2 Substitutional editing of mRNAs

#### 4.3.2.1 The occurrence and efficiency of mRNA editing

Eight of the ten mRNAs contain editing sites. In contrast to the even distribution of sites along the rRNA sequences, editing sites on mRNAs tend to be located near 5’ or 3’ ends: out of 27 sites, 23 were within about 100nt of either the 5’ or 3’ end. The *psbB* mRNA has two consecutive A-to-G editing sites in its 5’ UTR, but all 25 sites in the other 7 mRNAs are in the coding regions and result in 21 amino acid changes (Table 4.5). These changes usually result from a single editing site, although in *psaA* two editing sites together change Ile$^{731}$ to Ala.

The *petD* and *psbE* minicircles can produce polycistronic RNAs carrying both the protein-coding gene and three downstream tRNA genes. The transcripts are further processed to form the mature mRNAs tailed with a string of uridine (Nelson et al., 2007). We used genomic PCR, RT-PCR and 3’ RACE with different primer combinations to amplify *petD* partial sequences from DNAs, polycistronic RNAs and mature mRNAs, respectively (Figure 4.1). By comparing the sequencing electropherograms of the 3’ RACE products (mature mRNAs), genomic PCR (DNAs) and RT-PCR products (polycistronic RNAs), we observed that polycistronic RNA and DNA sequences are exactly the same, whereas mature RNAs contain a single A-to-G edit that converts Ile$^{145}$ to Val. Similar results were obtained from three other editing sites in *petD*, suggesting that substitutional editing should occur during or after the *petD* transcript is cleaved from the polycistronic RNA precursor and acquires a poly(U) tail.

In animals, RNA editing, especially A-to-I editing, can be easily recognized by searching for mixed A/G peaks on sequencing chromatograms, given that RT-PCR products are directly sequenced. However, there was no obvious overlapping peak at
the editing site of either precursors or mature mRNAs (Figure 4.1), suggesting that the RNA editing converts bases very efficiently and thoroughly.

4.3.2.2 Substitutional editing creates some canonical start codons

The analysis of *H. triquetra* minicircle DNAs indicated that 6 of 10 protein-coding genes (*psbB, psbC, psbE, psaA, psaB* and *atpA*) use AUA as a start codon (Zhang et al., 1999; Nelson et al., 2007). Only *psbD, petB* and *petD* have the canonical AUG start codon, while *psbA* uses an alternative prokaryotic start codon, UUG. Our 5′RACE data show that the putative AUA start codon of *psbC* and *psbE* is converted to an AUG start codon with A-to-G editing. However, *psbB, psaA* and *psaB* still keep AUA as the start codon. As all the 5′ RACE data represent the 5′ end of mature chloroplast mRNAs (see Methods), AUA is therefore an alternative start codon.

In the case of *atpA*, we found that an A-to-C editing located 26nt upstream of the AUA converted a stop codon into serine, which added another 27 amino acid residues to the previously predicted N-terminus (Figure 4.2). Although the N-terminus of the shorter ORF (open reading frame) matches those of *L. polyedrum* and *A. carterae*, we believe that the extended version of the ORF is more likely to represent the natural situation, because it leaves a 45nt 5′ UTR (untranslated region), which agrees with the average length (40nt) of the 5′ UTR of other chloroplast mRNAs.

4.3.2.3 Effect of editing on conservation of protein primary sequence

In order to evaluate the functional effects of editing, we aligned seven protein sequences predicted from edited mRNAs with homologs from dinoflagellates (*H. triquetra, Amphidinium carterae* and/or *Lingulodinium polyedrum*), other red line species (*Odontella sinensis, Thalassiosira pseudonana* and *Phaeodactylum tricornutum, Guillardia theta, Rhodomonas salina, Emiliania huxleyi* and *Cyanidioschyzon merolae*) and green line species (*Chlamydomonas reinhardii, Physcomitrella patens, Zea mays* and *Arabidopsis thaliana*). Most editing sites that result in amino acid changes probably have little effect on protein function: out of 19 changed amino acid residues, 13 have side chains with similar properties to those of unedited sequences (Table 4.5). Sixteen of the changes increased the identity to at least
some of the dinoflagellate homologs, including all six cases where the nature of the amino acid was changed by editing.

**4.3.3 16S rRNAs are in pieces**

Previous studies showed that dinoflagellate chloroplast 16S and 23S rRNAs are actually in pieces (Wang and Morse, 2006; Nisbet et al., 2008). When 5’ RACE was carried out with gene-specific primer 16SR4 (Figure 4.3A), there were two products, one of which was the mature 5’ end (Fragment 1) and the other the result of an internal cleavage at Site a (Fragment 2). Using 5’-RACE with primers 16Sr3, 16Sr2 and 16Sr1, three more internal cleavage sites were detected (Figure 4.3A, Sites b, c and d). Based on these results, the 16S rRNA coding region is 1514nt, excluding the poly(U) tail. The theoretical size for each fragment ranges from 135 to 538bp, but since the 3’ ends of Fragments 1-4 were not experimentally determined, we cannot rule out the possibility of further processing leading to smaller fragment sizes.

To confirm that 16S rRNA is indeed in pieces, we used two probes, one covering the first two fragments (P1) and the other covering the other three fragments (P2) to perform northern blotting (Figure 4.3B). The northern analysis for P1 showed a weak band of about 0.5kb and a thick band of about 0.2 kb. The former could be an unprocessed 472bp precursor and the intense band could contain both Fragment 1 (135bp) and Fragment 2 (337bp). Since only one intense band (0.2kb) was detected, Fragment 2 might be trimmed at the 3’ end. In the northern analysis with P2, two intense bands of about 0.2kb and 0.5kb as well as two weak bands (1.0kb and 1.2kb) were detected. It is possible that the 0.2kb band represents Fragment 4 and the thick 0.5kb band is made of both Fragment 3 and Fragment 5. The higher molecular mass bands are probably unprocessed precursors.

This fragmented RNA is not likely an artifact, as little smearing was observed. The 5’ RACE results showed that the four cleavage sites are reproducible. Note that the RNA blotting results of light-phase and dark-phase RNA samples were not significantly different, suggesting that the pattern of rRNA processing is not affected by diurnal
rhythm, though the signal of RNAs from the dark phase is slightly weaker than that from the light phase (Figure 4.3B).

### 4.3.4 A secondary structure model of the edited 16S rRNA

A total of 39 editing sites were found in 16S rRNA, of which 36 sites are A-to-G, two G-to-A and one U-to-A. The editing slightly increases the GC%, suggesting that it may increase the thermostability of the RNA molecules. However, as rRNAs interact with many proteins, the folding of rRNAs does not exactly conform to the laws of lowest free energy. Since the dinoflagellate 16S rRNA is highly divergent and hence poorly aligned with other homologs (Zhang et al., 2000), a better approach to understanding the possible functional and evolutionary importance of rRNA editing would be to rely on the higher-order structure.

We carried out the secondary structure analysis of *H. triquetra* 16S rRNA with the comparative RNA method (Cannone et al., 2002). The post-edited *H. triquetra* sequence was aligned with the *E. coli* 16S rRNA sequence to locate the highly conserved motifs (6 to 10nt) shared between them, which are usually at loops or bulges in terms of the secondary structure. These motifs served as markers to help locate the positions of less conserved regions, usually predicted as helical in the *E. coli* model. The analyses of these helical regions used the co-variation approach (Cannone et al. 2002) where the potential helix may not have the same composition but maintains G:C, A:U and G:U base-pairing. Together, regions with conserved motifs and predicted helices are called conserved regions in terms of their proper folding (highlighted on the *E. coli* structure, Figure 4.4 insert), although the primary sequence is poorly aligned. For those regions that are highly variable or contain large gaps, we used the MFOLD program (Zuker, 2003) to predict possible folding formats and chose the structures closest to those from *E. coli* 16S rRNA, because its three-dimensional structure is known (Wimberly et al. 2000). As this method relies heavily on sequence alignment, the structure of these regions is hypothetical. Finally the structure of the folding results from both regions were put together manually to produce the folding model of *H. triquetra* chloroplast 16S rRNA (Figure 4.4). In general, the *H. triquetra* 16S rRNA shows considerable structural resemblance to the *E. coli* model.
Surprisingly, the region carrying the anti-Shine-Dalgarno (anti-trigger) sequence is replaced by the poly(U) tail.

Although the structure of the variable regions might be arbitrary, many regions can form secondary structures that strictly conform to the model with exact helical length and similar loop sequences (highlighted, Figure 4.4 insert), making it possible to assess the functional and evolutionary effect of editing events within these regions. Twenty edited sites are located in conserved regions, including 19 A-to-G and one G-to-A change. Interestingly, only 3 editing sites are involved in creating G-C pairs. Others are either in loop regions or turn an A-U pair into a weak G•U pair.

To better assess the evolutionary importance, we compared the conserved regions from our model with both *E. coli* and a predicted folding model of *Guillardia theta* 16S rRNA (http://www.rna.ccbb.utexas.edu/). We found that 11 editing events restore conservation (Figure 4.4, arrow heads). An example is shown in Figure 4.5. In this conserved region (H23 and H23b), there are four editing sites (bold) and three of them (shaded) are changed to conform to the two other structures. The fourth change appears to be neutral, but it might still be useful to form a non-canonical A•G base pairing to maintain the structure. One edit changed an A-U pair to a weaker G•U pair, but it restored the conservation of the site. In the well conserved regions of Figure 4.4, there are 6 cases where an A-U pair is changed to a G•U pair. This indicates that the edited bases might have interactions with other nucleotides or ribosomal proteins to maintain higher order structure.

### 4.4 Discussion

The 16S rRNA is one of the major components of ribosomes and is, of course, essential for the expression of chloroplast genes. However, the chloroplast 16S rRNA sequence of *H. triquetra* is highly divergent and some workers have suggested it may be a pseudogene (Koumandou et al., 2004). However, if this were the case, it would raise the question of where the true 16S rRNA gene was located, given that organellar import of such a large RNA has never been observed in any organism. However, as a major component of the 30S ribosome, the function of 16S rRNA is determined by its
proper folding. To better understand this 16S rRNA, we deduced its secondary structure with an RNA comparative method. Although some divergent regions are hard to predict, conserved regions conform well to the *E. coli* model (Figure 4.4). Furthermore, there is even more similarity with a model based on 131 chloroplast 16S rRNA sequences, which can be downloaded from http://www.rna.ccbb.utexas.edu/ (data not shown).

The four major cleavage sites which break the rRNA into five experimentally detected fragments (Figure 4.3) are found in the variable regions (Figure 4.4). Fragmented rRNA has been widely reported in bacteria, where the cleavage sites are only found in the variable regions (Evguenieva-Hackenberg, 2005). Fragmented chloroplast rRNA was also observed in *Chlamydomonas eugametos* (Turmel et al., 1991). In dinoflagellate mitochondria, even the rRNA genes are fragmented (Jackson et al., 2007; Kamikawa et al., 2007).

We compared the mutation data from *E. coli* 16S rRNA to further evaluate our model (Yassin et al., 2005). We found that *H. triquetra* 16S rRNA contains almost all the essential bases where mutations are deleterious in *E. coli*. In addition, *H. triquetra* 16S rRNA also contains most of the essential bases that have been determined to be critical for the A site, the P site and translation accuracy (Yassin et al., 2005).

Chloroplast RNA editing occurs in the thecate dinoflagellates *H. triquetra*, *C. horridum* and *L. polyedrum*, but not in the athecate *A. carterae*. Here we showed that editing plays a role in maintaining protein function by increasing the conservation of coding regions among dinoflagellates. In addition, we found that several predicted AUA start codons were edited to canonical AUG codons, but that *psaA*, *psaB* and *psbB* still appear to use AUA as a start codon. Therefore, in total three types of start codons (AUG, AUA and UUG) are used for 10 protein coding genes, raising the question of why such a small genome needs so many different start codons. Since only three tRNA genes have so far been discovered on dinoflagellate chloroplast minicircles (Nelson et al. 2007), it is possible that a nucleus-encoded tRNA with a UAU anticodon is imported into the chloroplasts along with the other “missing” tRNAs.
The most intriguing question for dinoflagellate organellar RNA editing concerns the mechanisms. Among those substitutional editing types identified in dinoflagellates so far, A-to-G editing occurs most frequently. Since the animal A-to-I editing is read as A-to-G editing, Zauner et al (2004) argued that the A-to-G editing in dinoflagellate minicircle transcripts may take a similar biochemical pathway to modify adenosine to inosine. However, the A-to-I editing conserved in metazoans is catalyzed by ADARs (adenosine deaminase acting on RNA) which requires double-stranded RNA (>20bp) as the target (Bass, 2002). Such an RNA duplex is usually formed by base-pairing of a segment of RNA containing the editing sites with an adjacent complementary segment (editing site complementary sequence, ECS) located either in an exon or an intron (Jepson and Reenan, 2007). We ran a computational analysis with MFOLD, (Zuker, 2003) to try to locate the ECS element around the A-to-G editing sites in the *H. triquetra petD* mRNA. We found that in most predicted structures generated by the MFOLD program the A-to-G editing sites are in loops or the border of helices (data not shown), but we cannot draw any conclusion so far. Indeed, computational analysis for ECS elements requires confirmation by *in vitro* experiments, as the folding of mRNA can be very dynamic (Reenan, 2005). In contrast, rRNAs have relatively rigid folding patterns and this makes it reasonable to ask if the site-specific A-to-G editing in dinoflagellate chloroplast rRNA needs an RNA duplex as a target. In the conserved regions from our 16S rRNA folding model (Figure 4.4), most of the edited sites are either in the loop regions or helix bulges. Therefore, we hypothesize that at least the A-to-G editing in rRNAs does not use the same mechanism as animal A-to-I editing.

Apart from A-to-G editing, U-to-C editing accounts for 30% of editing sites in mRNAs. U-to-C editing is frequently observed in organellar transcripts of hornworts and ferns (Kugita et al., 2003; Wolf et al., 2004). Although little is known about U-to-C editing, some studies in mammals showed that C-to-U editing, the reverse process of U-to-C editing, is catalyzed by a cytidine deaminase (Smith, 2007). In plants such deaminase homologs were also identified (Hegeman et al., 2005). However, to date none of the cytidine deaminases was found to be able to catalyze the reverse reaction, an amination step leading to the U-to-C changes. So the mechanism for U-to-C editing, along with other types such as G-to-C editing, remains a mystery.
In summary, editing of *H. triquetra* chloroplast gene transcripts is dominated by A to G conversions and occurs at a late stage of RNA maturation. Two of the five AUA start codons deduced from genomic sequence were edited to canonical AUG codons, but the other three were not; thus confirming AUA as a legitimate start codon in this species. Comparative modeling of the edited 16S rRNA showed that it shares conserved secondary structural elements with other 16S rRNAs in spite of its very divergent primary sequence, supporting its role as a functional component of the chloroplast ribosome.

**Acknowledgements:**

We thank Dr. Claudio Slamovits for helpful directions in using the comparative RNA methods. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canada Council to BRG.
Table 4.1 Primers used for 5' RACE of *H. triquetra* chloroplast transcripts.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Purpose (detecting the 5' end of)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16Sr1</td>
<td>TTG ACG GGC AGT ATG TGT AAA G</td>
<td>16S rRNA Fragment 3</td>
</tr>
<tr>
<td>16Sr2</td>
<td>TCG GCT ATT CCT GAA CTG AAC C</td>
<td>16S rRNA Fragment 2</td>
</tr>
<tr>
<td>16Sr3</td>
<td>TCA CAC CTC AAG CTT TCG TGA</td>
<td>16S rRNA Fragment 1, 1st PCR</td>
</tr>
<tr>
<td>16Sr4</td>
<td>CAA CCA AAG GGT GAC GAA GT</td>
<td>16S rRNA Fragment 1, 2nd PCR</td>
</tr>
<tr>
<td>23Sr2</td>
<td>CGG GCT TCT CCC ATT TGT AAC A</td>
<td>23S rRNA</td>
</tr>
<tr>
<td>atpAr1</td>
<td>CCG ACG TCA ATA GCT GGC TTA</td>
<td>atpA, 1st PCR</td>
</tr>
<tr>
<td>atpAr2</td>
<td>GGA ACC ATT GCG TCA ACA GAT</td>
<td>atpA, 2nd PCR</td>
</tr>
<tr>
<td>petBr1</td>
<td>GAC CAA CAC TGA AGC CAC CAC</td>
<td>petB, 1st PCR</td>
</tr>
<tr>
<td>petBr2</td>
<td>ACC AGC ACC CGG AAG TAA ATC</td>
<td>petB, 2nd PCR</td>
</tr>
<tr>
<td>petDr1</td>
<td>CAT GAT TGG TCT TCT GAA TGG A</td>
<td>petD, 1st PCR</td>
</tr>
<tr>
<td>petDr2</td>
<td>CAA ACG GAT TAG CAG GTT CAC</td>
<td>petD, 2nd PCR</td>
</tr>
<tr>
<td>psaAr</td>
<td>ACC GAG ATC CTG GTT GAG AAT</td>
<td>psaA</td>
</tr>
<tr>
<td>psaBr</td>
<td>CCA CCT GAA GTA GCT GAA ACG</td>
<td>psaB</td>
</tr>
<tr>
<td>psbAr1</td>
<td>AGT GAA CCA CCG AAG ACA GCA</td>
<td>psbA, 1st PCR</td>
</tr>
<tr>
<td>psbAr2</td>
<td>GTG GAG CTG AGA AAG CAA CGA</td>
<td>psbA, 2nd PCR</td>
</tr>
<tr>
<td>psbBr</td>
<td>CAA TGC CAG AAA GCT GCT AAG</td>
<td>psbB</td>
</tr>
<tr>
<td>psbCr1</td>
<td>TAC CGA GAT GAG CAC CAA GAA T</td>
<td>psbC, 1st PCR</td>
</tr>
<tr>
<td>psbCr2</td>
<td>AAA CAC CGC AGA CGA AAT AAG A</td>
<td>psbC, 2nd PCR</td>
</tr>
<tr>
<td>psbDr</td>
<td>CCA TAA GAA CCC AAC GAG TGA</td>
<td>psbD</td>
</tr>
<tr>
<td>psbEr</td>
<td>TGG GTT GAA GTT TGG AAC ACC</td>
<td>psbE</td>
</tr>
<tr>
<td>5IP</td>
<td>CGC GGA TCC GAA CAC TGC GTT TGC TGG CTT TGA TG</td>
<td>5' end Inner Adaptor Primer, Ambion</td>
</tr>
<tr>
<td>5OP</td>
<td>GCT GAT GGC GAT GAA TGA ACA CTG</td>
<td>5' end Outer Adaptor Primer, Ambion</td>
</tr>
</tbody>
</table>
Table 4.2 Primers used in the 3’ RACE of *H. triquetra* chloroplast transcripts.

The Utail Adaptor was used for cDNA synthesis as described in Chapter 2. Unlike the 5’ RACE, one round of PCR with a gene specific primer and 3OP (3’ end outer primer, Invitrogen) was sufficient to produce satisfactory results.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>16sRNAf</td>
<td>GTG CAT GGC TGT TAA AGG GTA</td>
</tr>
<tr>
<td>atpAf</td>
<td>GCT GCT TAC AGT GGT GCT GCT</td>
</tr>
<tr>
<td>petBf</td>
<td>CCA AGG TGC TTC AGG TTT CG</td>
</tr>
<tr>
<td>petDf1</td>
<td>GCC TTA TCTCAA GAG TAC CAT GTT</td>
</tr>
<tr>
<td>psaAf</td>
<td>TGC TGG TAT TCC TAG CGC AAA</td>
</tr>
<tr>
<td>psaBf</td>
<td>TTC GCT TGT GAT GGT CCA GGT A</td>
</tr>
<tr>
<td>psbAf</td>
<td>TCG TTT TCT CAG CTC CAC</td>
</tr>
<tr>
<td>PsbBf</td>
<td>ATT CCG TTC TGG TCC ACA GCT C</td>
</tr>
<tr>
<td>psbCf</td>
<td>CTG GTG TCC AGG TGG TTG</td>
</tr>
<tr>
<td>psbDf</td>
<td>GTG GTG CTC TCC TCT CAT T</td>
</tr>
<tr>
<td>psbEf</td>
<td>GGC TAT AAG TAG AGG TGA ACG T</td>
</tr>
<tr>
<td>Utail Adaptor</td>
<td>GCG AGC ACA GAA TTA ATA CGA TGT ACT ATA GGA AAA AAA AAA AAA AAA AAA BN</td>
</tr>
<tr>
<td>3OP</td>
<td>GCG AGC ACA GAA TTA ATA CGA CT</td>
</tr>
</tbody>
</table>
Table 4.3 Primers for the RT-PCR of *H. triquetra* chloroplast transcripts to obtain the cDNA sequence data uncovered by the 5’ and 3’ RACE

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Primer combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>16SrRNAf1</td>
<td>AGT GCA AGG CAA TGA TGG GTA T</td>
<td>16SrRNAf1+16Sr3</td>
</tr>
<tr>
<td>16SrRNAf2</td>
<td>GGT GTG ACA GGG ATT AGA TAC</td>
<td>16SrRNAf2+16Sr2</td>
</tr>
<tr>
<td>atpINf</td>
<td>TGG TCG TGG TCA GCG TGA ATT A</td>
<td>atpINf+atpINr</td>
</tr>
<tr>
<td>atpINr</td>
<td>ATC CGC AGC GAA CTG AGA GAA T</td>
<td></td>
</tr>
<tr>
<td>psaIN1f</td>
<td>TGT TTT CTT CTG GCT CGG TGG T</td>
<td>psaIN1f+psaIN1r, for the 1st half of the psaA mRNA</td>
</tr>
<tr>
<td>psaIN1r</td>
<td>AAA GGC GTG AAT GCC ATG AGT T</td>
<td></td>
</tr>
<tr>
<td>psaIN2f</td>
<td>TGT TTT CTT CTG GCT CGG TGG T</td>
<td>psaIN2f+psaIN2r, for the 2nd half of the psaA mRNA</td>
</tr>
<tr>
<td>psaIN2r</td>
<td>AGT GAA ATG GGC ACC AAG GAA A</td>
<td></td>
</tr>
<tr>
<td>psaIN2f</td>
<td>CGC CAG AGT AAT GGT GCA GCT A</td>
<td>psaIN1f+psaIN1r, for the 1st half of the psaB mRNA</td>
</tr>
<tr>
<td>psaIN1r</td>
<td>CAC CAA CCG GAG AGA TTG TGT G</td>
<td></td>
</tr>
<tr>
<td>psaIN2f</td>
<td>TCA TTT CGG TGC TCA AGG TGA A</td>
<td>psaIN2f+psaIN2r, for the 2nd half of the psaB mRNA</td>
</tr>
<tr>
<td>psaIN2r</td>
<td>GGA AAA TCC ATG CCC AAA CTT G</td>
<td></td>
</tr>
<tr>
<td>psbINf</td>
<td>GGG AAG CTC TTG GTT TCG ATG A</td>
<td>psbINf+psbINr</td>
</tr>
<tr>
<td>psbINr</td>
<td>ATA ACT GGC CAA GCA GCA AGG A</td>
<td></td>
</tr>
<tr>
<td>psbINf</td>
<td>GTT GGG CAG CCG TTA TGA CTT T</td>
<td>psbINf+psbINr</td>
</tr>
<tr>
<td>psbINr</td>
<td>ATT CAG CAC GAC GGA ATG GAA T</td>
<td></td>
</tr>
<tr>
<td>PsbCINf</td>
<td>CAC ATT TCG CTC CTG AAA AGC</td>
<td>psbCINf+psbCINr</td>
</tr>
<tr>
<td>PsbCINr</td>
<td>AAC CAC CTT GGA CAG ACC AGA</td>
<td></td>
</tr>
<tr>
<td>psbDINf</td>
<td>TTA GCT GCT GGT GGT TGG TTC A</td>
<td>psbDINf+psbDINr</td>
</tr>
<tr>
<td>psbDINr</td>
<td>ATG CGA CAC CGA AAA TCT GTG A</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.4 Editing of \textit{H. triquetra} chloroplast RNA sequences determined by RACE and RT-PCR.

<table>
<thead>
<tr>
<th></th>
<th>Number of edited sites (%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mRNA\textsuperscript{a}</td>
<td>rRNA\textsuperscript{b}</td>
</tr>
<tr>
<td>A to G</td>
<td>14</td>
<td>53</td>
</tr>
<tr>
<td>G to A</td>
<td>0</td>
<td>4 (6.5%)</td>
</tr>
<tr>
<td>A to C</td>
<td>1 (3.7%)</td>
<td>0</td>
</tr>
<tr>
<td>U to C</td>
<td>7 (25.9%)</td>
<td>0</td>
</tr>
<tr>
<td>G to C</td>
<td>3 (11.1%)</td>
<td>0</td>
</tr>
<tr>
<td>A to U</td>
<td>1 (3.7%)</td>
<td>2 (3.3%)</td>
</tr>
<tr>
<td>U to A</td>
<td>0</td>
<td>2 (3.3%)</td>
</tr>
<tr>
<td>C to U</td>
<td>1 (3.7%)</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>61</td>
</tr>
</tbody>
</table>

% of sites edited

\textsuperscript{a} Data include the full mRNA sequence of \textit{psaA}, \textit{psbB}, \textit{psbC}, \textit{psbD}, \textit{psbE}, \textit{petB}, \textit{petD} and \textit{atpA}.

\textsuperscript{b} Data include the full sequence of 16S rRNA (1514nt) and the partial 5’ end sequence of 23S rRNA (675nt).
<table>
<thead>
<tr>
<th>Gene</th>
<th>Editing sites</th>
<th>Amino acid change</th>
<th>Consensus amino acid</th>
<th>Dinoflagellates</th>
<th>Red line</th>
<th>Green Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>atpA</td>
<td>Aaa-Gaa</td>
<td>(K6E)</td>
<td>I</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atpA</td>
<td>uAa-uCa</td>
<td>(stop-S18)</td>
<td>I</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atpA</td>
<td>auA-auG</td>
<td>(I28M)</td>
<td>Mf</td>
<td>I,V</td>
<td>I,V</td>
<td></td>
</tr>
<tr>
<td>atpA</td>
<td>Auc-Guc</td>
<td>(I31V)</td>
<td>V,T</td>
<td>I,V</td>
<td>/</td>
<td></td>
</tr>
<tr>
<td>atpA</td>
<td>gUu-gCu</td>
<td>(V88A)*</td>
<td>A,C</td>
<td>/</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>atpA</td>
<td>cAu-cGu</td>
<td>(H451R)</td>
<td>V</td>
<td>K</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>petB</td>
<td>gUu-gCu</td>
<td>(V127A)*</td>
<td>T,C</td>
<td>V</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>petD</td>
<td>Aua-Gua</td>
<td>(I41V)*</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>petD</td>
<td>Auu-Guu</td>
<td>(I145V)*</td>
<td>V,Y</td>
<td>I</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>petD</td>
<td>Gcu-Ccu</td>
<td>(A150P)*</td>
<td>P,G</td>
<td>P</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>petD</td>
<td>aUu-aCu</td>
<td>(I152T)</td>
<td>T,S</td>
<td>/</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>petD</td>
<td>Aaa-Gaa</td>
<td>(K153E)</td>
<td>E</td>
<td>K,Q,I,D</td>
<td>K,I</td>
<td></td>
</tr>
<tr>
<td>psaA</td>
<td>Uuu-Cuu</td>
<td>(F4L)</td>
<td>R,Y</td>
<td>S</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>psaA</td>
<td>guA-guU</td>
<td>(V249V)*</td>
<td>/</td>
<td>P</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>psaA</td>
<td>uaU-uaC</td>
<td>(Y599Y)*</td>
<td>/</td>
<td>I,V</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>psaA</td>
<td>AUu-GCu</td>
<td>(I731A)</td>
<td>L,A</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>psbC</td>
<td>auA-auG</td>
<td>(I1Mf)</td>
<td>Mf,V</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>psbC</td>
<td>Auu-Guu</td>
<td>(I3V)</td>
<td>V</td>
<td>F</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>psbC</td>
<td>gaA-gaG</td>
<td>(E335E)*</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>psbC</td>
<td>Ggu-Cgu</td>
<td>(G444R*)</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>psbD</td>
<td>gUu-gCu</td>
<td>(V46A)*</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>psbD</td>
<td>Uuu-Cuu</td>
<td>(F149L*)</td>
<td>I,L</td>
<td>I</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>psbE</td>
<td>auA-auG</td>
<td>(I1M)</td>
<td>Mf</td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>psbE</td>
<td>aGa-aCa</td>
<td>(R3T)*</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

**Key**

“/”: the editing site is located in a variable region of the alignment.

“-”: the editing site is located in a gap in the alignment.

“*”: the editing site is located in a long conserved block
Figure 4.1 Substitutional editing of the petD minicircle transcripts.

The arrangement of petD and downstream tRNA genes in the minicircle is shown at the top. Arrows indicate the approximate location of each primer. Total RNA was used to make cDNA using random primers (cDNA (random)) or the Utail Adaptor (cDNA (3’RACE)). The 3OP primer (Ambion) used for 3’RACE is complementary to Utail Adaptor. Primers are listed in Table 4.2 and Table 2.1. The editing site is highlighted in the sequencing chromatograms.
Figure 4.2 Editing makes a new stop codon for atpA. The mRNA sequence starts at the transcriptional start site determined by 5’ RACE.

The editing sites are marked with asterisks. The second of the two boxed AUG’s is the start codon predicted based on the DNA sequence (Zhang et al., 1999). The shaded letters show the stop codon that is converted to a serine codon by A-to-C editing.
Panel A: Four cleavage points (a, b, c, d) separate the 16S rRNA into five fragments. The cleavage sites were detected by 5’ RACE with gene-specific primers 16Sr4, 16Sr3, 16Sr2 and 16Sr1, respectively, using cDNA prepared with random primers. Panel B: RNA blotting of *H. triquetra* total RNA with the two 16S rRNA probes (P1 and P2) shown as long arrows in A. RNA was purified from cells harvested at the mid-light phase (L) or mid-dark phase (D) and separated on a 1% formaldehyde agarose gel. The Riboruler™ RNA Ladder (Fermentas) was used as size marker.
The structure was predicted based on the structure of *E. coli* 16S rRNA (insert) using the comparative method. Bold letters indicate the sites edited in *H. triquetra*. Unless noted, the type of editing is A-to-G change. Arrow heads indicate that the edited sites increase the identity to the models (*E. coli* and *G. theta*), available from the Comparative RNA web site (Cannone et al., 2002). The five fragments detected by 5’ RACE (Figure 3) and their 5’ and 3’ ends are labelled; roman numerals indicate the three conventional domains. The helical numbering (e.g. H44 for Helix 44) follows the definition in Wimberley et al (2000). The highlighted regions of the insert are structures strictly conserved between *E. coli* and *H. triquetra*, whereas the other parts are variable regions. Analysis for editing effects was only applied to the conserved regions.
Figure 4.5 Comparison of partial secondary structures among *H. triquetra*, *G. theta* and *E. coli*.

A region conserved among all taxa (helix H23 and H23b) is enclosed in the box. Bold letters indicate the post-edited sites, all of which are A-to-G substitutions, and those that restore the conservation of the sites are highlighted. The corresponding sites in *G. theta* and *E. coli* are also highlighted.
4.5 References


Chapter 5 Possible loss of eubacterial-like chloroplast RNA polymerase and cloning of a bacteriophage-like RNA polymerase*

*A version of this chapter will be submitted for publication

5.1 Introduction

Dinoflagellates are important marine algae. They are not only one of the major primary producers in marine ecosystems, but also a serious threat to the ecosystem, fishery and human health when they form blooms. Typically, photosynthetic dinoflagellates have secondary chloroplasts bounded by three membranes. Accordingly, the protein targeting to the chloroplast requires at least two pieces of information at the leader sequence: ER signal sequence and transit sequence (Nassoury et al., 2003, Patron et al., 2005) Dinoflagellate chloroplasts possess a unique genome comprised of a group of minicircles of usually 2-3kb. In most cases one minicircle carries one gene. So far only 17 genes have been found on minicircles, in contrast to a conventional chloroplast genome carrying 100-200 genes on a master circle (reviewed in Howe et al., 2008).

Typically, chloroplast genes are transcribed with a multisubunit RNA polymerase inherited from the ancestral cyanobacterial endosymbiont. This eubacteria-like RNA polymerase has 5 subunits (α, β, β’, β” and σ, encoded by, rpoB, rpoC1, rpoC2 and rpoD, respectively) and is found in plastids of all plants and algae (Smith and Purton, 2002). Since the genes of core components (rpoA, rpoB, rpoC1, rpoC2) are usually located on chloroplast genomes, this enzyme is commonly called PEP (plastid-encoded RNA polymerase). It recognizes the canonical -10/-35 element similar to that of eubacterial σ70-type promoter (Allison, 2000). Chloroplasts of higher plants have an additional bacteriophage-like RNA polymerase, which is thought to have been derived from mitochondrial counterpart by gene duplication and retargeting one copy to the chloroplast (Smith and Purton, 2002). This RNA polymerase only has only one subunit (encoded by rpoT) and is called NEP (nucleus-encoded RNA polymerase) as its gene is located in the nucleus. It is typically responsible for the transcription of housekeeping genes (Smith and Purton, 2002). However, it is not always the case. In some parasitic plants, the rpo genes encoding the PEP have been lost from the plastid genome. Thus the whole plastid genome is transcribed by the NEP (Funk et al., 2007). NEP recognizes a different type of promoter characterized with a YRT motif, as well
as a GAA box located at ~10-20 bp upstream of the YRT motif in many cases (Weihe and Borner, 1999).

Proper function of dinoflagellate chloroplasts needs an RNA polymerase to transcribe their minicircle genes. However, none of the minicircles discovered so far encodes any gene of the PEP. The dinoflagellate chloroplast genome is severely reduced due to massive lateral gene transfer from the chloroplast to the nucleus (Bachvaroff et al., 2004; Hackett et al., 2004). The fact that many genes highly conserved in chloroplast genomes of plants and algae have been relocated to the nucleus in dinoflagellates, raises the question of whether the genes encoding the PEP have been relocated to the nucleus. In this study, we combined PCR, Southern, northern and western analyses to provide preliminary evidence supporting the conclusion that PEP has been lost in dinoflagellates. We also cloned a \textit{rpoT} gene, which might be used for minicircle transcription.

\textbf{5.2 Materials and methods}

\textbf{5.2.1 Northern blotting and Southern blotting}

An axenic culture of \textit{Heterocapsa triquetra} (CCMP 449) was obtained from the Provasoli–Guillard Culture Center for Marine Phytoplankton (Boothbay Harbor, ME) and grown in f/2–Si media at 18° with 12-h light/12-h dark cycles at 50 μmol m\textsuperscript{-2} s\textsuperscript{-1} light intensity. Cells were usually collected at 28 days after inoculation (cell density approximately 10\textsuperscript{6} cell/mL), corresponding to the early stationary phase of growth.

The methods for RNA extraction, northern blotting and radioisotope-labelled probe preparation were described in Dang and Green (2009) (Chapter 4). Total DNA extraction and southern blotting followed the methods described in Zhang et al. (1999), except that DNA was not purified using CsCl-Hoechst dye density gradient centrifugation. Blots were hybridized overnight at 45 °C and washed at 40 °C.

\textbf{5.2.2 Primers}

The CODEHOP primers for \textit{rpoB} were designed with the web tools provided by the authors (http://bioinformatics.weizmann.ac.il/blocks/codehop.html). The degenerate
primers for \textit{rpoT} were adapted from Cermakian et al. (1996). All primers and their sequences are listed in Table 5.1.

5.2.3 Inhibitor (rifampicin) test

An axenic culture of \textit{H. triquetra} (1 L) during late stage of logarithmic growth (2 weeks after inoculation) was divided into 4 aliquots, which were treated with 0 µM, 50 µM, 100 µM and 200 µM rifampicin (dissolved in DMSO), respectively. The cultures were kept in dark for 8hr and used for RNA extraction. Exactly 500 ng of total RNA for each sample was used for the synthesis of first-strand cDNA, using Superscript III (Invitrogen) and the downstream primer psbAinr (Table 1). PCR reactions were carried out for 25 cycles: 94ºC for 30 s, 56ºC for 30 s followed by 72ºC for 1 min. 1 µL of cDNA from each sample was used in the 25 µL PCR reaction mixture containing 0.1 mM dNTP, 1×PCR buffer, 40 pmol primer, 1 mM MgCl₂, and 1 unit \textit{Taq} polymerase (Sigma).

5.2.4 Cloning of \textit{H. triquetra rpoT}

About 1µg of total RNA was used for cDNA synthesis primed by random hexamers (Invitrogen). The first-round PCR was carried out with rpoTdf1 and rpoTdr for 35 cycles: 94ºC for 30 s, 53ºC for 1 min, 56ºC for 20 s, 60ºC for 5 s, followed by 72ºC for 2 min. A 0.7 kb band was recovered from the gel and used as template for the next round of PCR (diluted 100 times). Second round PCR used rpoTdf2 and rpoTdr with the same conditions as the first round PCR to confirm that this 0.7kb band contains a \textit{rpoT} fragment. This 0.7kb band was then cloned and sequenced. Based on the sequence five gene-specific primers were designed (rpoTinvf1-3, rpoTinvr1-2) to obtain the 3’ end and partial 5’ end sequence by using a cDNA-based inverse PCR method (Huang et al., 2000). Combined with the sequence from the 0.7kb band, we obtained 2.1kb \textit{rpoT} sequence with this method, which contains the complete 3’ end of the \textit{rpoT} ORF and incomplete 5’ end sequence. Three gene specific primers (rpoTr 1-3) were designed based on the newly obtained 5’ end sequence to perform 5’ RACE (see methods in Chapter 4), which gave a 0.6kb product. Another three gene-specific primers (rpoTr4-6) were designed based on the sequence of the 0.6kb band to perform
5' RACE, which gave a 0.7kb product. All sequences were aligned to produce a 3kb contig.

5.2.5 Phylogenetic analysis
The protein sequence of \textit{rpoT} homologs, obtained from Genebank, were aligned with the \textit{rpoT} of \textit{H. triquetra} with Clustal X (version 1.83) and edited manually. Tree construction was carried out with Phylip (version 3.69) by using maximum likelihood method (ProML). The tree was evaluated with the same method (ProML) and 1000 replicates.

5.2.6 Western blotting
200mL cell cultures (\textit{H. triquetra}, \textit{Thalassiosira pseudonana}, \textit{Heterosigma akashiwo}) were spun down, resuspended with 5 mL extraction buffer (0.1M HEPES (pH7.2), 50mM KCl, 10mM EDTA, 0.626M Sorbitol, 1×protease inhibitor cocktail, 0.1%DTT), and disrupted by mini bead beater (Biospec) at full speed for 1 min with 0.1mm diameter beads. The lysate was centrifuged at 10,000g for 5 min and the supernatant was collected. Preparation of \textit{Zea mays} total proteins used the same extraction buffer but the tissue was ground with liquid nitrogen. Protein samples were separated with SDS-PAGE gels, transferred electrophoretically to nitrocellulose membranes (Amersham), and blocked with 5% milk in Tris-buffered saline containing 0.05% Tween 20 for more than 1hr. The blots were first incubated with 1:1000 dilution of anti-ZmrpoTp (gift of Dr. D.B. Stern) or a 1:500 dilution of anti-rpoB (gift of Dr. M.R. Hansen) and then with a 1:10,000 dilution of commercial peroxidase-linked secondary antibodies. Signals were detected with the chemiluminescence system (Amersham).

5.3 Results
5.3.1 Degenerate PCR could not amplify \textit{rpoB} from \textit{H. triquetra} total DNA
Since PEPs are universally encoded in the plastid genomes of photosynthetic organisms but not found in the minicircle of any dinoflagellate species so far, we postulated that dinoflagellates might retain this set of genes in the nucleus by lateral gene transfer. Although none of the PEP genes has been identified from previous
studies on the expressed sequence tag (EST) profiles (Bachvaroff et al., 2004; Hackett et al., 2004; Patron and Keeling, 2005), more EST data released in recent years prompted us to re-examine the existence of PEP genes. Given that dinoflagellate chloroplast genes are very divergent, we chose the \textit{rpo} gene sequences from apicomplexan \textit{Plasmodium} plastid as the baits to search for the homologs from GenBank and the taxonomically broad EST database (TBestDB, \url{http://tbestdb.bcm.umontreal.ca/}) by using tBlastn. We did not obtain any hit, suggesting that PEP genes might have been lost rather than migrating to the nucleus. However, it is also possible that PEP genes are not actively transcribed or that the PEP genes still reside in the chloroplasts. So we decided to search for PEP genes with PCR and Southern blotting.

We chose \textit{rpoB} gene (coding $\beta$ subunit) as our target for database searching as it is more conserved than the rest of the subunits (data not shown). In an attempt to amplify \textit{rpoB}, we turned to a primer design method termed CODEHOP (consensus-degenerate hybrid oligonucleotide primers), which is widely used in amplifying weakly conserved genes with PCR at high efficiency (Rose et al., 1998). Different from traditional degenerate primers, CODEHOPs only allow degeneration at 10-12 bases of the 3’ end of the primer. The advantage of this primer design over traditional methods is that PCR will not deplete one specific primer within the pool after many cycles. As the two primers (\textit{rpoBdnf} and \textit{rpoBdnr}) were designed based on the C terminal conserved motifs of \textit{rpoB}, they should be able to amplify \textit{rpoB} from a wide spectrum of species. As shown in Figure 5.1A, genomic PCR with CODEHOP primers (\textit{rpoBdnf} and \textit{rpoBdnr}) gave a 0.5-0.7kb band from \textit{Arabidopsis thaliana} (gift from Dr J.-G. Chen), \textit{Thalassiosira pseudonana} (gift from Dr. S. Zhu), \textit{Cyanophora paradoxa} and \textit{Escherichia coli}. PCR products were subsequently sequenced, confirming that this pair of primers is effective in amplifying \textit{rpoB} regardless of the GC content of the target genes (GC\% of \textit{rpoB} in \textit{E. coli} is 68\% while in \textit{T. pseudonana} is 37\%). Therefore, even though the GC\% of nuclear and chloroplast genes varies considerably in \textit{H. triquetra} (65\% in nucleus and 36\% in chloroplasts), the primers should be able to amplify \textit{rpoB} regardless of gene location. However, the \textit{H. triquetra} sample did not give any products with the estimated length. We also used this pair of primers for
Amphidinium carterae and Adenoides eludens, but no positive result was obtained (data not shown).

5.3.2 Southern analysis gave no signal of rpoB gene

Since negative PCR results cannot be regarded as solid evidence of rpoB gene loss, we performed Southern blotting using the rpoB fragments of T. pseudonana as a probe because diatoms and dinoflagellates are related based on phylogenetics. Since the dinoflagellate nuclear genome is heavily methylated (ten Lohuis and Miller, 1998), we selected three restriction enzymes (BamHI, KpnI and PstI) insensitive to methylation so that the genomic DNA could be completely digested. A low-stringent hybridization condition was used to allow the probe to cross-react with a heterologous rpoT. The Southern analysis results are shown in Fig 5.1B. We observed no signal from genomic DNA of H. triquetra, suggesting that the rpoB might have been lost. However, the probe (rpoB of T. pseudonana) did not strongly react with A. thaliana rpoB PCR product (from Lane AT in Fig 1A). It is still possible that a divergent rpoB gene may present in H. triquetra since the DNA sequence could be highly variable even though the protein sequence is conserved.

5.3.3 Rifampicin did not affect the expression of psbA in H. triquetra

Rifampicin is a widely used antibiotic that specifically inhibits the transcription initiation of eubacterial-like RNA polymerase while the bacteriophage-like RNA polymerase remains unaffected. Rifampicin showed a marked inhibitory effect on the plastid transcription of Plasmodium, a relative of dinoflagellates (Lin et al., 2002). We used rifampicin to test for the presence of a conventional PEP in dinoflagellate plastids. However, as shown in Figure 1C, similar to the control (GAPDH1, nucleus-encoded), no significant change of psbA transcripts was detected with the increase of rifampicin concentration from 0 – 200μM. This could indicate the loss of PEP in A. carterae plastids. However, this may also have resulted from the natural resistance of dinoflagellates to rifampicin (Smith and Purton, 2002), or rapid degradation of rifampicin in seawater.
5.3.4 Tobacco RpoB antibody gave no signal for *H. triquetra* total protein extract

Since RpoB is highly conserved at the protein level, we used an antibody raised against the tobacco RpoB for western analysis (Hegeman et al., 2005). As shown in Figure 5.1D, the samples produced bands larger than 110kD. Based on the prediction of the ORFs, the size of the RpoB protein from *T. pseudonana*, *H. akashiwo*, and *Z. mays* is 150kD, 125kD and 120kD, respectively. Therefore, the western analyses suggest that this antibody against tobacco RpoB is able to recognize heterologous RpoB proteins. We detect no strong signal from *H. triquetra* protein extract, suggesting that the RpoB might either have been lost or become highly divergent.

5.3.5 Cloning dinoflagellate *rpoT*

As we obtained no evidence to support the existence of *rpoB* in *H. triquetra*, we reasoned that the minicircle transcription could be carried out by a bacteriophage-like RNA polymerase (RpoT), as in the chloroplast of some parasitic plants (Funk et al., 2007). By using degenerate PCR, inverse PCR and 5’ RACE (see methods), we obtained a *rpoT* mRNA with a total length of 3052bp. The ORF is 2865bp, which translates into a protein of 954 amino acids (106kD). We used this sequence to construct a phylogenetic tree. As shown in Figure 5.2, the *rpoT* of *H. triquetra*, together with the early branching dinoflagellate *Perkinsus*, is closely related to that of apicomplexans. This suggests that the cloned *H. triquetra rpoT* is indeed a dinoflagellate *rpoT* sequence.

To identify the copy number of *rpoT* in *H. triquetra*, we performed Southern analyses for *H. triquetra* by using a 0.7kb PCR product of *rpoT* as a probe. As the PCR product is located in the highly conserved C-terminus of the *rpoT*, the probe should be able to recognize other homologous *rpoT* genes at low stringent conditions. However, although five bands can be recognized from PstI-digested DNA sample, only one band can be recognized from BamHI and KpnI-digested samples, suggesting that only one copy of the *rpoT* gene is present in the *H. triquetra* genome (Figure 5.3A).

We analyzed the first 100 amino acid residues at the RpoT N-terminus and found no leader sequence that would guide it to either chloroplasts or mitochondria. So we
performed a northern analysis by using a 0.7kb PCR product (corresponding to partial 3’ end of \textit{rpoT}) as a probe to analyze the expression of \textit{rpoT} (Fig 5.3B). A faint band of about 3kb can be identified from the RNA samples collected in both the light and dark phase, suggesting that the cloned \textit{rpoT} cDNA sequence should be very close to the full size of the \textit{rpoT} mRNA. To further characterize the size of \textit{H. triquetra rpoT}, we carried out western blotting by using an antibody against maize rpoTp, which was raised with a 100 a.a. peptide at the C terminal of rpoTp (Chang et al., 1999). As the C-terminus of RpoT is highly conserved across a wide spectrum (Cermakian et al., 1997), the antibody can cross react with \textit{in vitro}-expressed \textit{H. triquetra rpoT}. (data not shown). We used this antibody to analyze the size of \textit{in vivo}-expressed \textit{H. triquetra} RpoT and found that the size is about 100kD, which suggests that the mRNA sequence we obtained is at least close to the full length of \textit{rpoT} (Figure 5.3C).

\textbf{5.4 Discussion}
Although there is no doubt that minicircles are the chloroplast genome and actively transcribed (Howe et al., 2008), the transcriptional machinery of dinoflagellate minicircles remains as a mystery. As most of the chloroplast genomes are transcribed by a bacteria-like multi-subunit RNA polymerase, we first aimed at detecting the indispensible β subunit (encoded by \textit{rpoB}) to verify the presence of bacteria-like RNA polymerase in dinoflagellate \textit{H. triquetra}. However, the degenerate PCR, Southern western analyses gave negative results. Although none of the negative results can be regarded as hard evidence of \textit{rpoB} gene loss, so far no report claims to find minicircles carrying \textit{rpoA}, \textit{B} or \textit{C}. Moreover, none of the \textit{rpo} homologs can be identified in the released EST data for a variety of dinoflagellate species. Therefore, we hypothesize that \textit{rpoB} might have been lost.

Since we have no evidence to support the existence of bacteria-like RNA polymerase, we turn to an alternative candidate gene \textit{rpoT} which encodes a single-subunit RNA polymerase (RpoT). This polymerase is an ideal candidate for minicircle transcription for several reasons. First, RpoT does not need any auxiliary co-factor to initiate transcription (Hedtke et al., 2000). In this case transcription initiation is less sequence dependent, but it agrees with the evidence that minicircle transcription might be
randomly started in *H. triquetra* (Chapter 2). Second, as RpoT only has one subunit and is much smaller in size than the multisubunit RNA polymerase (i.e. PEP), it can quickly respond to signals from chloroplasts. Third, the advantage of PEP is that the nucleus can control specific operons by expressing and importing σ factor (RpoD) into chloroplasts. This is useful for a conventional chloroplast genome carrying 100-200 genes. However, dinoflagellate chloroplasts do not require such a complicated transcriptional control mechanism as most chloroplast genes have moved to the nucleus.

We encountered many technical obstacles when cloning the *rpoT* gene, mainly due to the large genome and low mRNA level (Fig 5.3B). Indeed, for a certain minicircle, the copy number is only about 10 during the exponential phase of cell growth (Koumandou and Howe, 2007). Since *H. triquetra* only has 17 species of minicircles (*psbA, psbB, psbC, psbD, psbE, psaA, psaB, petB, petD, atpA, 16S rRNA, 23S rRNA* minicircles and five aberrant minicircles), an actively growing cell might only have about 100-200 minicircles. Therefore, the dinoflagellate chloroplasts do not demand high expression level of *rpoT* during their growth, making it hard to clone the *rpoT* gene.

For the 3 kb *rpoT* sequence we obtained, we detected no poly(A) tail at the 3’ end with the inverse RT-PCR method (Huang et al., 2000). Indeed, our attempt to obtain complete 3’ end UTR by using 3’ RACE also gave no results. Moreover, we found it is hard to amplify a partial sequence of *rpoT* with gene specific primers and using poly(T)-primed cDNA as templates. These results suggest that this mRNA might not have a poly(A) tail. The 5’ end of *rpoT* is another mystery. The ORF predicted from the 3kb sequence gives the protein size of 106kD, which is close to the actual size of RpoT detected by western analysis (about 100kD, Figure 5.3C). However, analyzing the N-terminal leader sequence indicates that it contains no signal sequence, nor mitochondrial targeting sequence (data not shown). Therefore, it is still likely that the 3kb sequence is only a part of the *rpoT* mRNA. We continued the 5’ RACE and inverse RT-PCR by using new primers designed based on the 3kb *rpoT* sequence, but the sequence of the PCR products showed the same 5’ end as it was previously
determined. Most dinoflagellate mRNAs are post-transcriptionally capped with a 22nt spliced leader sequence at the 5’ end (Lee et al., 2007; Zhang et al., 2007). We used the spliced leader sequence as the forward primer and \textit{rpoT} gene-specific reverse primer to perform RT-PCR but no band was produced.

In summary, in this study we used PCR, northern and Southern and western analysis to acquire preliminary data of possible RNA polymerase governing the minicircle transcription. We hypothesize that dinoflagellate chloroplasts have abandoned the multisubunit bacterial-like RNA polymerase and use a bacteriophage-like RNA polymerase to transcribe minicircle genes.
Table 5.1 The primers used in semi-quantitative RT-PCR and cloning \textit{rpoB} and \textit{rpoT}

<table>
<thead>
<tr>
<th></th>
<th>Degenerate PCR for \textit{rpoB}</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{rpoBdnr}</td>
<td>CCT CCA GGG CCC ACA YYT CCA TYT C</td>
</tr>
<tr>
<td>\textit{rpoBdnf}</td>
<td>GTG CGG CCG GCA YGG NMA YAA RG</td>
</tr>
<tr>
<td>Inhibitor tests</td>
<td></td>
</tr>
<tr>
<td>\textit{GAPDH1f}</td>
<td>GAC TCC CTG GTG ATC GAT GG</td>
</tr>
<tr>
<td>\textit{GAPDH1r}</td>
<td>ACC GTG TGC GAG TCA TCC TT</td>
</tr>
<tr>
<td>\textit{psbARf}</td>
<td>CAT GGG TCG TGA ATG GGA AT</td>
</tr>
<tr>
<td>\textit{psbARr}</td>
<td>TGA ACC ACC GAA GAC AGG</td>
</tr>
<tr>
<td>Degenerate RT-PCR for \textit{rpoT}</td>
<td></td>
</tr>
<tr>
<td>\textit{rpoTdf1}</td>
<td>GITSITGCAACGGIYTNCARCA</td>
</tr>
<tr>
<td>\textit{rpoTdf2}</td>
<td>GGTIGTIAAGCAGACNGTNATGAC</td>
</tr>
<tr>
<td>\textit{rpoTdr}</td>
<td>GCGTIGTCCARWAISWRTCRTG</td>
</tr>
<tr>
<td>Inverse RT-PCR for \textit{rpoT}</td>
<td></td>
</tr>
<tr>
<td>\textit{rpoTinvr1}</td>
<td>CCA GCT GCC GCG TTT CTA CGA</td>
</tr>
<tr>
<td>\textit{rpoTinvr2}</td>
<td>GGC CTC CCT CTC AAC CTT CTC</td>
</tr>
<tr>
<td>\textit{rpoTinvr3}</td>
<td>GTC GCT GGG CGT GAG GGT CA</td>
</tr>
<tr>
<td>\textit{rpoTinvr1}</td>
<td>GGT GGA ACT GCC TTG GAC G</td>
</tr>
<tr>
<td>\textit{rpoTinvr2}</td>
<td>ATC GCC GTC GCA CCT TCT</td>
</tr>
<tr>
<td>5’ RACE for \textit{rpoT}</td>
<td></td>
</tr>
<tr>
<td>\textit{AB5 OUP}</td>
<td>GCT GAT GGC GAT TAA TGA ACA CTG</td>
</tr>
<tr>
<td>\textit{AB5 INP}</td>
<td>CGC GGA TCC GAA CAC TGG TGG TGG CTT TGA TG</td>
</tr>
<tr>
<td>\textit{rpoTr1}</td>
<td>ACA GGA TTG GCC TCC CTC TCA AC</td>
</tr>
<tr>
<td>\textit{rpoTr2}</td>
<td>AGA CCA TTG CAG GTT CCG TCA AG</td>
</tr>
<tr>
<td>\textit{rpoTr3}</td>
<td>GAC TCG CGC ATC AAA GGT GAT CT</td>
</tr>
<tr>
<td>\textit{rpoTr4}</td>
<td>GAG CAC ACAG GAT CTC CTC</td>
</tr>
<tr>
<td>\textit{rpoTr5}</td>
<td>AGC ATG TGG TCC AGC ACC GTA</td>
</tr>
<tr>
<td>\textit{rpoTr6}</td>
<td>CTT CAC TTC GTT GTC CGC ATC</td>
</tr>
</tbody>
</table>
Figure 5.1 The evidence supporting the loss of rpoB in *H. triquetra*

A: genomic PCR to amplify *rpoB* with CODEHOPs (rpoBdnf and rpoBdnr). About 50ng of genomic DNA was used for each reaction. B: Northern analysis of *H. triquetra*. 15ng of genomic DNA was digested overnight with BamHI (B), KpnI (K) and PstI (P). 5pg of *rpoB* PCR products from *A. thaliana* and *T. pseudonana* (Lane AT and TP, respectively) was used as a positive control to test the cross reaction. The probe was prepared with the PCR product of *T. pseudonana* rpoB (Lane TP, fig 5.1A). The blot was hybridized with probes at 45° and washed at 40°. C: Inhibitor (rifampicin) test. Four aliquots of *H. triquetra* cell culture were treated with rifampicin of 0, 50, 100µM and 200µM for 8 hrs. For each sample 0.5µg of total RNA was used for the semi-quantitative RT-PCR. D: Western analysis of RpoB. For each lane, 20µg of total protein extracts from each sample species (HT: *H. triquetra*, TP: *T. pseudonana*, HA: *H. akashiwo*, ZM: *Zea mays*) were loaded and antibody against tobacco RopB (Hegeman et al., 2005) was used to detect RpoB.
Figure 5.2 Phylogenetic analysis of *H. triquetra* *rpoT*.

The tree was constructed and evaluated with the maximum likelihood method-based program (ProML) provided in Phylip (version 3.69). Only the bootstrap supports larger than 50% are shown at the node.
Figure 5.3 The copy number and expression of *H. triquetra* *rpoT*

A: Southern analysis. The blot was prepared as described in Fig 1B. The probe was prepared with a 0.7 kb PCR product corresponding to the conserved 3' end of *rpoT*. B. Northern analysis. Each lane contains 25 µg of total RNA extracted from 15-day cell culture at light phase (L, i.e. 6hr after light on) or dark phase (D, i.e. 6hr after light off). The probe is the same as in A. C: Western analysis: 20µg total proteins of *Zea mays* and *H. triquetra* were separated with 8% SDS-PAGE. An antibody against maize RpoTp was used in this analysis.
5.5 References


Chapter 6 Conclusions and future directions
6.1 Summary of major discoveries of this thesis

A major goal of my PhD study was to investigate the mechanisms of minicircle transcription and post-transcriptional processing of minicircle RNAs in dinoflagellate *H. triquetra*. Before I started this project, it had been established that minicircles are the genome of dinoflagellate chloroplasts and the minicircle transcripts are extensively edited in some dinoflagellate species. Moreover, 3’ end polyuridylation had also been discovered. My studies extended knowledge of dinoflagellate minicircle gene expression mainly in two aspects: 1) for the first time put forward a “rolling-circle” transcription mechanism; 2) analyzed the possible roles and mechanisms of RNA editing, especially in the 16S ribosomal RNA. The major discoveries are summarized below:

1. The tRNA genes located in the *petD* and *psbE* minicircle are co-transcribed with the upstream protein-coding gene. This discovery indicated that the transcription of minicircles first produces a large polycistronic pre-RNA, which is then processed into discrete mature RNAs. (Chapter 2)

2. The minicircles of *H. triquetra* use a unique rolling-circle transcription mechanism. Transcription is likely randomly initiated and produces long transcripts. Northern and RT-PCR analyses suggest that some long transcripts carry two repeats of the template sequence (Chapter 3).

3. The mature 3’ end of both mRNAs and tRNAs is processed from the long transcripts by a single-step endonucleolytic cleavage (Chapter 3). Although the 3’ end processing mechanism for tRNAs appears similar to those in plant mitochondria, the mechanism for mRNAs has never been found in any other organisms.

4. Extensive substitutional RNA editing was found in the mature mRNA but not in the precursor RNAs, suggesting that the editing of chloroplast RNA occurs concomitantly with RNA processing (Chapter 4).

5. The northern analyses suggest that the chloroplast 16S rRNA is in pieces with the 5’ RACE method. Four internal cleavage sites were revealed, which break the 16S rRNA into five fragments (Chapter 4).

6. Using a bioinformatic approach I produced the first 16S rRNA secondary structure model of dinoflagellate chloroplasts. The model suggests that the *H. triquetra* chloroplast 16S rRNA could be functional even though the sequence is highly divergent. Moreover, the
model suggests that A-to-G editing, the most abundant type of substitutional editing in both mRNA and rRNA, might not occur by the same mechanism as in animals (Chapter 4).

7 The rpoB gene encoding the β subunit of the PEP core enzyme might have been lost in H. triqueta as the gene and its product cannot be detected with degenerate PCR, Southern blotted, inhibitor test and western blotting. I cloned the first dinoflagellate NEP gene (rpoT), but it remains to be determined where this protein is targeted to (Chapter 5).

6.2 The transcription initiation of minicircle and the single subunit RNA polymerase

In the “rolling-circle” transcription model, one of the questions remaining to be answered is where the transcription initiation site (TIS) is. Since my attempt to detect the TIS was largely unsuccessful, I hypothesized that the transcription is randomly initiated (Chapter 3). This hypothesis, though somewhat arbitrary, could be partially supported by the possible loss of the PEP (bacterial-like RNA polymerase) in H. triqueta (Chapter 5) in that the PEP core enzyme cannot initiate transcription without the promoter-recognition factor (i.e. sigma factor, RpoD) (Allison, 2000). On the other hand, the NEP (bacteriophage-like RNA polymerase) is able to initiate transcription without the co-factor. The best example is the yeast mitochondrial RNA polymerase (Rpo41). Typically, the yeast mitochondria RNA polymerase requires Mtf1 factor to guarantee the precise initiation (Shadel and Clayton, 1995; Jan et al., 1999). However, the function of Mtf1 factor is to facilitate melting the templates rather than recognize the promoter sequence. Therefore, Rpo41 could initiate transcription without the mtf1 factor if a bubble (>6 bp) is formed at the promoter region (Matsunaga and Jaehning, 2004). In the non-coding regions of H. triqueta minicircles, the low GC% (~30%) could facilitate the formation of bubbles throughout, providing many potential sites for the putative NEP to initiate transcription.

I cloned an rpoT from H. triqueta, which could serve as a candidate for minicircle transcription. However, no targeting signal can be recognized from the amino-terminus of the predicted ORF, suggesting that the rpoT sequence could be incomplete. I reasoned that this could result from the harsh disruption of the cell by using glass beads because the method would inevitably produce foam. To acquire to the full sequence, a better approach to gently smash the cell needs to be developed. If
the putative full sequence can be acquired, it would not only provide clear information of its targeting, but also could be used in *in vitro* tests for minicircle transcription if it indeed targets to the chloroplasts.

### 6.3 Crosstalk between RNA editing and RNA processing

As an inosine is recognized as a guanosine by the splicing machinery, in animal nucleus, A-to-I editing could create or delete splicing sites by modifying the 5’ splice donor site (from AU to I(G)U) or the 3’ splice acceptor site (i.e. from AA to Al(G)) (Athanasiadis et al., 2004; Valente and Nishikura, 2005). Extensive RNA editing has been found in dinoflagellate mitochondria and chloroplasts (Lin et al., 2002; Zauner et al., 2004). In *H. triquetra*, most editing events are found in the 3’ and 5’ end of mRNAs (Chapter 4). Although some editing events were shown to elevate the identity of proteins compared to their homologs, most appear to be neutral (Lin et al., 2002; Zauner et al., 2004). In an early study on tRNA processing plant mitochondria, Kunzmann et al. (1998) found that the RNase Z, an endonuclease processing the 3’ end of pre-tRNA, prefers edited tRNAs as the substrate, suggesting that RNA editing may serve as a signal for cleavage. In *H. triquetra*, RNA editing is rarely detected in precursor RNAs, suggesting that the cleavage might be controlled by RNA editing. This might partially explain why no recognizable high MW band can be found in the *psaB* lane (Fig 3.2): since *psaB* has no editing events, RNA precursors are rapidly cleaved without any inhibition.

### 6.4 Precise determination of the secondary structure of 16S rRNA

16S rRNA is indispensable for the translation machinery of chloroplast genes as it is the RNA component of the 30S ribosome. In *H. triquetra* chloroplast, this minicircle gene is highly divergent and was long believed to be a pseudogene (Koumandou et al., 2004). As the function of ribosomal RNAs depends on the three-dimensional structures, I predicted the *H. triquetra* 16S rRNA structure with a comparative method (Cannone et al., 2002). The structure generated by this method provides strong evidence that the *H. triquetra* 16S rRNA is likely to be functional (Chapter 4, Figure 4.4) in that it shares all conserved functional motifs with the *E. coli* model. However, this computation-based structure might be inaccurate in those variable regions as no reference is available to compare (e.g. Fragment 1, 2, Figure 4.4).
Directly determining the secondary structure of large molecules has long been a challenge. Even a relatively simple approach requires several chemical modification reagents to react with different types of nucleotides and separation of the reactants in polyacrylamide gels (Moazed et al., 1986). Due to the large expense and technique hurdles involved in such an approach, most of the rRNA folding models so far available have been generated by the comparative method (http://www.rna.ccbb.utexas.edu/). Recently, a relatively simple RNA structure determination approach, SHAPE (selective 2’-hydroxyl acylation analyzed by primer extension) was described (Merino et al., 2005). This method requires only one chemical reagent and uses many of the same tools as DNA sequencing for reactant separation and signal detection, which makes it possible to determine the structure of large molecules at low costs (Wilkinson et al., 2008; Watts et al., 2009). In the future, this approach could be used in accurately determining the H. triquetra 16S rRNA and other highly divergent RNA structures.
6.5 References


