Kính dâng hương hồn Ba!

Sequencing fragments of cryptophyte plastomes from 16S rRNA to *rbc*L genes and phylogenetic analyses based on the protein-encoding genes located in these fragments.

INAUGURAL-DISSERTATION
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Hoang-Dung Tran
aus Hochiminh City, Vietnam

Köln, 2009

Referees/Berichterstatter: Prof. Dr. Michael Melkonian
P.D Dr. Kerstin Hoef-Emden

Date of oral examination: October 27, 2009
(Tag der mündlichen Prüfung)

The present research work was carried out under the supervision of Prof. Dr. Michael Melkonian in the Botanical Institute, University of Cologne, Gyrhofstr. 15, 50931, Cologne, Germany from May 2003 to June 2006.

Diese Arbeit wurde von May 2003 bis Juni 2006 am, Botanisches Institut der Universität zu Köln, Gyrhofstr. 15, 50931 Köln, Germany unter der Leitung von Prof. Dr. Michael Melkonian durchgeführt.

ACKNOWLEDGEMENTS

This thesis was formed during my 4 year working at Lab of Prof. Dr. Melkonian lab, therefore, the first person I would like to thank is Prof. Dr. Melkonian, who was so kind and generous to accept me as a lab member when I thought I had no chance to be a student of *International Graduate School in Genetics and Functional Genomics*. He continuingly encouraged me when things was not as expectation and waited patiently for my completion of this thesis.

As one of the direct instructors, PD Kerstin Hoef-Emden gave me the very first lessons about phylogeny, algae and those precious techniques in the lab. Dr. Hoef-Emden also taught me the lifelong lessons of German working style, disciplined and enthusiastic.

Dr. Birger Marin guided me as friendly as an old brother, who I could rely on for experience and help.

I want to express my thanks to *International Graduate School in Genetics and Functional Genomics* for provided me the opportunity to be its student in 2002 and supported me to complete this thesis; and to the officers in the *International Graduate School in Genetics and Functional Genomics* for helping in various ways to clarify the things related to my academic works in time with excellent cooperation and guidance.

Thank you my friends, Jing Shi, Eva, Nelli, … and all other lab members that helped me in many unnamed things and shared all memories during my work here.

Thanks for the Ph.D thesis committee members paid attention and approved my thesis.

Last but not least, thank you my family and my beloved wife, who always stand by my side in all difficult situations.

LIST OF PUBLICATION

Part of this thesis have been published

Hoef-Emden, K., Tran, H.-D. & Melkonian, M. (2005) Lineage-specific variations of congruent evolution among DNA sequences from three genomes, and relaxed selective constraints on *rbc*L in *Cryptomonas* (Cryptophyceae). *BMC Evolutionary Biology* **5**:56.

ABSTRACT

In recent phylogenetic analyses combining nuclear and nucleomorph RNA genes of the ribosomal operons, three different colourless lineages were found in the genus *Cryptomonas*. This raised questions about the evolutionary history of these interesting objects and their relatives as well as the role of plastid genomes such as whether these three lineages resulted from similar or from different evolutionary events or what are the mutual relationship or/and roles of photosynthetic genes in the absence of photosynthetic activities, *etc*. To answer the interesting questions, the biological information has to been collected systematically from their plastid genomes.

At the first stage of the thesis, the cryptophyte plastid *rbc*L gene (1,5-biphosphate carboxylase/oxygenase [RuBisCO] large subunit) was chosen to amplify by BioTherm™ Taq DNA Polymerase and read their DNA compositions by SequiTherm EXCEL™ II DNA Sequencing Kit-LC and Li-Cor 4200L bidirectional sequencer. Eighteen newly *rbc*L sequences of *Cryptomonas* strains were obtained. Of these, five sequences were from heterotrophic (colorless) strains and the remaining was from photosynthetic (pigmented) strains. The results of *rbc*L phylogeny analyses showed that the colorless *C. paramecium* and their closely relative photosynthetic *Cryptomonas* had increased their evolutionary rates significantly. These were congruent with those of nuclear rDNA (concatenated SSU rDNA, ITS2 and partial LSU rDNA) and nucleomorph SSU rDNA that had been examined in previously. They were combined with other result done by Dr. Kerstin Hoef-Emden such as analyzing the shift from NNC to NNU in two-fold degenerate NNY codon in *rbc*L gene in *Cryptomonas* to discuss some hypotheses of the loss of photosynthetic activities in the colorless *C. paramaecium* strains. Detail results and discussion were published in BMC *Evolutionary Biology* 2005; **5**:56.

In the second part of the thesis, the goals were to amplify the cryptophyte plastome 16S rRNA-*rbc*L fragments by MasterAmp^{TM} Extra-long PCR kit and read their DNA sequences by BigDye Terminator v1.1 Cycle sequencing kit and automated ABI3730 sequencer, then exploited the sequencing information for further understanding the evolutionary history of cryptophyte plastomes. The task also attempted to find new evidence to explain the relationship between the changing from autotrophic to heterotrophic lifestyle in colorless *Cryptomonas* lineages and the elevation of evolutionary rates of photosynthetic genes that were located in the plastome 16S rRNA-*rbc*L fragment.

Twenty-two cryptophyte strains (four of them were colorless) were participated in this part. Most of the fragments (15) were read completely as planned while several fragments (7) were not, due to lack of time. The colorless strains possessed the smallest fragments; their plastomes, thus, were predicted to be the smallest among those of *Cryptomonas*. Strain *C. erosa* CCAC 0018 and *C. obovoidea* CCAC 0031 seemed to have the largest plastomes as their 16S-*rbc*L fragments contained an additional gene – *ycf*26 – that was not found in other *Cryptomonas* strains. Advantages and disadvantages of long-range PCR and primer-walking sequencing combination were discussed.

Based on the conserved domain analyses, all *ycf*26 from secondary plastids seems to be inactive and on the way to become pseudogene than alter its function. Another additional gene – ORF403 encoding Tic22 protein – also was examined the conserved domains and done a phylogenetic analysis. Some specific characteristics of ORF403 in rhodoplasts and cryptophyte plastome were found.

Three protein-coding genes – *chl*I, *rps*4 and *rbc*L – were used as separated phylogenetic markers or in combined. The results confirmed that one colorless lineage (presented by CCAC 0056, CCAP 977/2a, M2452, M2180) had accelerated evolutionary rates in all gene or/and protein trees.

The observations also suggested that *chl*I gene increased its substitution rate earlier than *rps*4 and *rbc*L genes as well as the elevated evolutionary rates could be ordered by *chl*I > *rps*4 > *rbc*L. Although having moderate size (609 bp), *rps*4 had an evolution rate neither as high as in *chl*I gene nor as low as in *rbc*L gene, producing acceptable phylogenetic trees for both nucleotide and protein levels. Therefore, *rps*4 gene seems to be more suitable protein-enoding plastid gene maker for phylogenies than its sisters, *chl*I and *rbc*L genes.

The ratio of NNC to NNU in two-fold degenerate NNY codons was calculated for each gene and discussed. It is possible that the shift in codon usage from NNC to NNU did not correlate to the relaxation of functional constraints and/or reduction of gene expression levels. Furthermore, the usage of NNU codons over the NNC in two-fold degenerate NNY codon seemed to be controlled by neutral mutation pressure rather than by selection followed by the gradually acceleration of evolutionary rate

A hypothetical scenario for the relations among the loss of photosynthesis, increasing of substitution rate of interring genes and time of diverging in colorless lineages was discussed.

Zusammenfassung

In den letzten phylogenetische Analysen mit nuklearen und nucleomorph RNA-Gene der ribosomalen Operons wurden drei farblos verschiedenen Linienin der Gattung *Cryptomonas* gefunden. Diese Fragen über die evolutionäre Geschichte dieser interessanten Objekten und deren Angehörige sowie die Rolle der Plastiden Genome, ob diese drei Linien aus ähnlichen oder aus verschiedenen evolutionären Ereignisse oder sind die gegenseitigen Beziehungen und / oder Rollen der Photosynthese-Gene in das Fehlen der photosynthetischen Aktivitäten usw. Um diese interessanten Fragen zu antworten werden die biologischen Daten systematisch aus ihren Plastiden Genome gesammelt.

In der ersten Phase des Thema wurde rbcL Gene in den cryptophyten Plastiden (1,5-biphosphate Carboxylase / Oxygenase [RuBisCo] großen Untereinheit) zur Ergänzung von Biotherm™ Taq DNA Polymerase gewaehlt und zum Lesen der DNA-Kompositionen von SequiTherm EXCEL™ II DNA-Sequenzierung Kit-LC und Li-Cor 4200L. Achtzehn von rbcL Sequenzen *Cryptomonas* wurden gesammelt. Fünf davon Sequenzen wurden aus heterotrophen (farblos) und die übrigen Teile wurde von photosynthetischen (pigmentiert) gesammelt. Die Ergebnisse der rbcL Phylogenie Analysen zeigten, dass die farblose *C. paramecium* und ihre eng relative photosynthetische *Cryptomonas* ihre evolutionäre Entwicklung deutlich hatten. Diese waren deckungsgleich mit denen der nuklearen rDNA (verketteten SSU rDNA, ITS2 und teilweise LSU rDNA) und nucleomorph SSU rDNA, die in zuvor geprüft worden. Sie wurden zusammen mit anderen Ergebnis von Dr. Kerstin Hoef-Emden, wie die Analyse der Verschiebung von NNC zu NNU in zwei-fach entartet NNY Codons in *rbc*L Gen in *Cryptomonas* zu diskutieren einige Hypothesen der Verlust der photosynthetischen Aktivitäten in den farblos *C. paramaecium* Stämme. Detaillierte Ergebnisse und Diskussionen wurden in BMC *Evolutionary Biology* 2005, **5**:56 veroeffentlicht.

Im zweiten Teil der Doktorarbeit sind die Ziele zur Ergänzung der cryptophyten plastome 16S rRNA-rbcL Fragmente von MasterAmp™ Extra-lange PCR Kit und lesen die DNA-Sequenzen von BigDye Terminator v1.1 Cycle Sequencing-Kit und automatisierte ABI3730 Sequenzer, dann nutzte die Sequenzierung Informationen für das Verständnis der Evolutionsgeschichte der cryptophyten plastomes. Die Aufgabe auch versucht, neue Beweismittel zu erklären, das Verhältnis zwischen dem Wechsel von

autotrophen zu heterotrophen Lebensweise in farblose *Cryptomonas*-Linien und die Höhe der Evolution von photosynthetische Gene, die in der plastome 16S rRNA-rbcL Fragment sind.

Zweiundzwanzig cryptophyte Stämme (vier von ihnen wurden farblos) waren an diesem Teil. Die meisten der Fragmente (15) wurden vollständig lesen wie geplant, während mehrere Fragmente (7) nicht wegen Mangel an Zeit wurden. Die farblose Stämme im Besitz der kleinsten Fragmente, deren plastomes, so wurden vorhergesagt, werden die kleinsten unter den von *Cryptomonas*. Stamm *C. erosa emend* CCAC 0018 und *C. obovoidea emend* CCAC 0031 zu haben schien die größte plastomes als 16S rRNA-rbcL Fragmente enthalten ein zusätzliches Gen - *ycf*26 - das war nicht in anderen *Cryptomonas* Stämme. Vor-und Nachteile der Langstrecken-und PCR-Primer-Walking-Sequenzierung Kombination erörtert.

Auf der Grundlage der erhaltenen Domain-Analysen scheinen alle *ycf*26 von sekundären Plastiden inaktiv zu sein und auf dem Weg zu pseudogene als ihre Funktion verändern. Eine weitere zusätzliche Gen - ORF403 Codierung Tic22 Protein - auch wurde die konservierte Domänen und phylogenetische Analyse geleistet. Einige Besonderheiten der ORF403 in rhodoplasts und cryptophyten plastome wurden gefunden.

Drei Protein-kodierenden Gene – *chl*I, *rps*4 und *rbc*L - wurden als phylogenetische Marker getrennt oder in Kombination. Die Ergebnisse bestätigten, dass ein farbloses Linie (von CCAC 0056, CCAP 977/2a, M2452, M2180) beschleunigten evolutionären in allen Gen-und / oder Protein-Bäume hatte.

Die Beobachtungen auch darauf hin, dass *chl*I Gen erhöhte seine Substitution Rate früher als rps4 und rbcL Gene sowie die erhöhten Sätze nach *chl*I> *rps*4> *rbc*L evolutionären könnte. Obwohl mit mäßiger Größe (609 bp), *rps*4 eine Entwicklung Rate weder so hoch wie in *chl*I -Gen noch so niedrig wie in *rbc*L Gen produziert akzeptabel phylogenetische Bäume für beide Nucleotid-und Protein-Ebene war. Daher *rps*4 Gen scheint besser geeignet enoding Plastiden-Protein-Gen-Maker für phylogenies als seine Schwestern, Chli und *rbc*L Gene zu sein.

Das Verhältnis von NNC auf NNU in zwei-fach entartet NNY Codons wurde für jedes Gen und diskutiert. Es ist möglich, dass die Verschiebung der Codon-Nutzung von NNC zu NNU korrelierte nicht zur Entspannung der funktionellen Einschränkungen und / oder Verringerung der Genexpression Ebenen. Außerdem ist die Nutzung von NNU Codons über die NNC in zwei-fach entartet NNY Codon durch neutrale Mutation Druck und nicht durch Auswahl, gefolgt von der schrittweise Beschleunigung der evolutionären Kurs

Ein mögliches Szenario für die Beziehungen zwischen den Verlust der Photosynthese, der zunehmenden Substitution von interring Gene und Uhrzeit der unterschiedlichen Linien war in farblos.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS 1
LIST OF PUBLICATION 2
ABSTRACT 3
TABLE OF CONTENTS 6
LISTE OF TABLES 11
LIST OF FIGURES 12
LIST OF ABBREVIATIONS 14

1.INTRODUCTION 17
1.1 The principal characteristics of the Cryptophyta cell 18
1.2 The genomes of cryptophytes 22
1.3 Sequencing strategies for Cryptophyta genomes 24
1.4 Aims of study 29

2.MATERIALS AND METHODS 31
2.1 Chemicals and medium 32
2.2 Algal culture medium 33
2.3 Harvest the cells for DNA isolation 38
2.4 Isolation of total genomic DNA by CTAB method 38
2.5 Amplify the plastid *rbc*L gene of *Cryptomonas* strains by BioTherm™ Taq DNA Polymerase and read their DNA compositions by SequiTherm EXCEL™ II DNA Sequencing Kit-LC and Li-Cor 4200L bidirectional sequencer 41
2.5.1 Construct primers *41*
2.5.2 Set up conditions for PCR running *44*
2.5.3 Purify PCR products by use streptavidin-coated dynabeads M-280 system *45*
2.5.4 Prepare Pre-Mix and Mix for PCR-based sequencing *45*
2.5.5 Read the nucleotide compositions *46*
2.5.6 Proofread newly sequenced rbc*L gene* *47*
2.5.7 Deposit the sequence data *47*

2.6 Amplify the the cryptophyte plastome 16S rRNA-*rbc*L fragments by MasterAmpTM Extra-long PCR kit and read their DNA sequences by BigDye Terminator v1.1 Cycle sequencing kit and automated ABI3730 sequencer 47
2.6.1 Design the primers for long-PCR 48
2.6.2 Optimize the long-range PCR protocol and screen the primers ... 51
2.6.3 Verify the PCR products 52
2.6.4 Amplify the plastid 23S rRNA–rbcL fragment by optimal long-range PCR protocol 53
2.6.5 Produce the smaller segment by BioTherm™ Taq DNA Polymerase 53
2.6.6 Sequence the short PCR products (4kb) by primer-walking approach and using automated ABI3730 sequencer 56

2.7 Annotate the genes 64

2.8 Search for homologous protein by SMART server 64

2.9 Align based on codon 64

2.10 Analyze the phylogenetic issues 64
2.10.1 For nucleotide data sets 65
2.10.2 For amino acid sequences 65

3.RESULTS 67

3.1 Search and sequence the Cryptophyta plastome *rbc*L gene 68

3.2 Phylogenetic analyses based one the cryptophyte plastid *rbc*L gene... 70

3.3 Annotate the 16S rRNA–*rbc*L fragments 72
3.3.1 Gene contents 72
3.3.2 Overlapping genes 79
3.3.3 Comparison the lengths of the fragments 79

3.4 AT content and Identity characteristics 83
3.4.1 AT content 83
3.4.2 Identity characteristics 83

3.5 Detect Shine-Dalgarno sequence, -35 and -10 boxes of genes in 16S rRNA–*rbc*L fragments 88

3.6 *ycf*26 – Unexpected protein in *C. erosa* CCAC 0018 and *C. obovoidea* CCAC 0031 96
3.6.1 The distribution of ycf26 *in photosynthetic organisms 96*
3.6.2 Describe the functional structure of ycf26 *protein 97*
3.6.3 Search the functional domains of red-line plastid ycf26 *........ 98*

3.6.4 Reconstruct the phylogenetic trees ... *100*

3.7 ORF403 ... 108

3.7.1 Verify the translated ORF403 of cryptophytes as Tic22 protein. *108*

3.7.2 The distribution of Tic22 in host genomes ... *108*

3.7.3 Reconstruct the phylogenetic trees ... *110*

3.7.4 Search for the conserved domains ... *112*

3.7.5 Classify the Tic22 proteins ... *112*

3.8 Cluster of *psa*M and *chl*I gene and phylogenetic analyses based one the cryptophyte plastid *chl*I gene ... 121

3.8.6 psa*M and* chl*I as a conserved cluster in red-line plastid genomes* ... *121*

3.8.7 Annotate the conserved domains of chl*I protein* ... *121*

3.8.8 Reconstruct phylogenetic trees ... *122*

3.8.9 Codon usage ... *125*

3.9 Analyses based one the cryptophyte plastid *rps*4 gene ... 134

3.9.1 Annotate conserved domains ... *134*

3.9.2 Reconstruct the phylogenetic trees ... *134*

3.9.3 Codon usage ... *136*

3.10 Analyses based on the concatenated data sets of *chl*I, *rps*4 and *rbc*L genes ... 143

3.10.1 Reconstruct phylogenetic trees ... *143*

3.10.2 Calculate distance genetics ... *144*

4.DISCUSSIONS ... 148

4.1 Advantages and disadvantages of long-range PCR and primer-walking sequencing combination ... 149

4.2 *psa*M and *chl*I formed a cluster in red-line plastids. ... 151

4.3 The first evidence of reduced size of cryptophyte leukoplast genomes ... 152

4.4 First overviews lock about ORF403 ... 153

4.5 *ycf*26 in cryptophyte plastids seems to be a pseudogene than alters function ... 153

4.6 The evolutionary pathway of *chl*I seems to difference with those of SSU, LSU, *rps*4 and *rbc*L ... 156

4.7 The colorless group (M245, CCAC0056, CCAP 977/2a) diverged early or late depended on the gene surveyed. ... 157

4.8 Hypothetical scenario for the evolutionary history of colorless *Cryptomonas* 157

4.9 The shift from NNC to NNT in two-fold degenerate NNY codon seems not relative to the relaxation of the functional constraints of plastome protein-coding genes. 158

4.10 The usage of NNU codons over the NNC in two-fold degenerate NNY codon seems to be controlled by neutral mutation pressure rather than by selection followed by the gradually acceleration of evolutionary rate.... 159

5.PERSPECTIVES 160

6.APPENDIX 163

Appendix 1: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Rhodomonas* sp. M1480 plastome 164

Appendix 2: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Teleaulax* sp. SCAP K-416 plastome........ 164

Appendix 3: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. commutata* CCAC 0109 plastome........ 164

Appendix 4: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. loricata* M 2088 plastome........ 164

Appendix 5: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. erosa* CCAC 0018 plastome 164

Appendix 6: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. loricata* M 2088 plastome........ 164

Appendix 7: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. curvata* CCAC 0006 plastome 165

Appendix 8: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. borealis* CCAC 0113 plastome 165

Appendix 9: Illustration the gene content, gene order of the *rps*4 – *rbc*L fragment in *C. gyropyrenoidosa* sp. *nov*. CCAC 0108 plastome........ 165

Appendix 10: Illustration the gene content, gene order of the *rps*4–*rbc*L fragment in *C. lundii* CCAC 0107 plastome........ 165

Appendix 11: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. ovata* CCAC 0064 plastome 165

Appendix 12: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *C. paramaecium* CCAP 977/2a plastome........ 166

Appendix 13: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. paramaecium* CCAC 0056 plastome 166

Appendix 14: Illustration the gene content, gene order of the *chl*I–*rps*4 fragment in *C.* sp. M1634 plastome 166

Appendix 15: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment *in C. phaseolus* SAG 2013 plastome 166

Appendix 16: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. pyrenoidifera* CCMP 0152 plastome 166

Appendix 17: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. tetrapyrenoidosa* M 1092plastome 166

Appendix 18: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Chroomonas* sp. SAG 980-1 plastome............................167

Appendix 19: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Hemiselmis tepida* CCMP 0443 plastome.......................167

Appendix 20: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Proteomonas* sp. CCMP 0704 plastome 167

Appendix 22: Accession numbers of Cyanobaterial *ycf*26; AT% of the Cyanobacterial genome and *ycf*26 genes..169

7.REFERENCES 171

DECLARATION 186

CURRICULUM VITAE..187

LEBENSLAUF..189

LISTE OF TABLES

Table 2.2.1: List of Cryptophyta strains were used in this study 36

Table 2.5.1.1: PCR and sequencing primers for amplifying and sequencing cryptophyte *rbc*L gene 43

Table 2.6.1.1: Primers for long-range PCR to amplify the cryptophyte plastid 23S rRNA-*rbc*L fragments 49

Table 2.6.6.1: The universal sequencing primers for cryptophyte plastomes 16S rRNA-*rbc*L fragments 58

Table 2.6.6.2: The specific sequencing primers for some Cryptophyta strains to read the plastomes 16S rRNA-*rbc*L fragments 59

Table 3.1.1: List of *Cryptomonas* strains were read *rbc*L sequence 69

Table 3.2.1: Gene sizes in the cryptophyte plastome 16S rRNA–*rbc*L fragments . 80

Table 3.4.1: AT content of genes located between 16S rRNA and *rbc*L genes in examined Cryptophyta strains 85

Table 3.4.2: Comparison of identity values in non-photosynthetic and photosynthetic groups at both nucleotide and protein databases of genes in the Cryptophyta plastome 16S rRNA–*rbc*L fragments 87

Table 3.6.1: The distribution of *ycf*26 in plastomes 101

Table 3.6.2**:** The positions and E-values of functional domains for cryptophyte *ycf*26 protein 102

Table 3.7.1: E-values supported by SMART server for identify the Cryptophyta Tic22 proteins. 114

Table 3.7.2: The number and location of Tic22 genes in algae and land plant genomes. 115

Table 3.8.1: Base composition and variability of *chl*I sequences 126

Table 3.8.2: Codon usage of RNN, YNN and NNY codon in cryptophyte *chl*I gene. 127

Table 3.9.1: Base composition and variability of *rps*4 sequences 137

Table 3.9.2: Codon usage of NNC and NNY codon in cryptophyte *rps*4 gene 138

LIST OF FIGURES

Fig. 1.1.1: *Morphology of the cryptophytes. © 2008 Kerstin Hoef-Emden.*......21

Fig. 1.2.1: *Illustration of the secondary endosymbiotic process and four phylogenetically distinct genomes in the cryptophytes*....................................22

Fig. 2.5.1.1: *Approximate positions of the PCR sequencing primers.*...............42

Fig. 2.6.1.1: *Illustration the approximately positions, manner combinations of long PCR primers and the expected long PCR products using* G. theta *plastome as reference.*..50

Fig. 3.2.1*: Rooted maximum-likelihood tree inferred from* rbc*L gene sequences (51 taxa of 972 positions).*..71

Fig. 3.3.1: *Illustration the gene content, gene order of the 16S* rRNA*–rbc*L *fragment in* C. marssonii *CCAC 0086 plastome.* 76

Fig. 3.3.2: *Illustration the gene contents, gene orders of the 16S* rRNA*–rbc*L *fragment in* C. paramaecium *M2452 plastome.* 77

Fig. 3.3.3: *Illustration the gene content, gene order of the 16S* rRNA*–rbc*L *fragment in* C. o*obovoidea CCAC 0031 plastome.* 78

Fig. 3.5.1: *Alignment for 16S rRNA upstream region, one unknown-function conserved was found.*..90

Fig. 3.5.2: *The non-coding region of* psa*M gene,*..91

Fig. 3.5.3: *The upstream part of* chl*I gene,*..92

Fig. 3.5.4: *Alignment for upstream region of* rps*4 gene,* 93

Fig. 3.5.5*: The non-coding region of* rbc*L gene*..95

Fig. 3.6.1: *Alignment for* ycf26 *proteins of cryptophytes and some random chosen Cyanobacteria.*..103

Fig. 3.6.2*: Rooted maximum-likelihood tree inferred from 59* ycf26 *protein sequences (154 characters)*..106

Fig. 3.6.3: *Rooted maximum-likelihood tree reconstructed based on* ycf26 *gene sequences (59 taxa of 462 positions).*..107

Fig. 3.7.1: *Rooted maximum-likelihood tree reconstructed based on 68 ORF403 protein sequences (102 characters).* 116

Fig. 3.7.2: *Rooted maximum-likelihood tree inferred from ORF403 gene sequences (68 taxa of 306 positions).* 117

Fig. 3.7.3: *Alignment of Tic22 protein sequences originated from red-line plastids.* 118

Fig. 3.8.1: *The Genome ProtMap (NCBI) searching results showed the* psa*M*-chl*I clusters in some red-lineage algae* 128

Fig. 3.8.2: *The annotating for the conserved regions and hydrophilic and hydrophobic regions of* chl*I protein.* 129

Fig. 3.8.3: *Rooted maximum-likelihood tree inferred from* chl*I gene sequences (24 taxa of 963 positions).* 132

Fig. 3.8.4: *Rooted maximum-likelihood tree reconstructed based on 24* chl*I protein sequences (321 characters)* 133

Fig. 3.9.1: *Codon-alignment of* rps*4 protein sequences from cryptophytes, Cyanobacterium and some red-line plastids showed the numbers and order of α-helices and β-sheets.* 139

Fig. 3.9.2: *Rooted maximum-likelihood tree inferred from* rps*4 gene sequences (27 taxa of 597 positions).* 141

Fig. 3.9.3: *Rooted maximum-likelihood tree reconstructed based on 27* rps*4 protein sequences (199 characters).* 142

Fig. 3.10.1: *Rooted maximum-likelihood tree inferred from concatenated plastid* chl*I,* rps*4 and* rbc*L gene sequences (24 taxa of 2532 positions).* 145

Fig. 3.10.2: *Rooted maximum-likelihood tree reconstructed based on a concatenation of* chl*I,* rps*4 and* rbc*L protein sequences (24 taxa of 844 characters).* 146

Fig. 3.10.3: *The maximum likelihood distances (genetic distances) among cryptophytes with three data sets, strain* C. pyrenoidifera *CCMP 0152 was used as reference* 147

LIST OF ABBREVIATIONS

µM	micromolar
16S rRNA	16S ribosomal RNA
23S rRNA	23S ribosomal RNA
ADP	adenosine 5'-diphosphate
ATP	adenosine 5'-triphosphate
ATPase	adenosine 5'-triphosphatase
BAC library	Bacterial artificial chromosome library
bp	Base pair.
cDNA	complement DNA
***chl*I**	magnesium chelatase subunit I gene
CHM	chilomonas paramaecium medium
CTAB	Cetyltrimethylammonium Bromide
Da	Dalton
dATP	dexoxyadenosin triphosphate
dH_2O	distiller water
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	deoxynuclotide
dsDNA	double strand deoxyribonucleic acid
EDTA	ethlylendiamine tetraacetic acid
EMBL	European Molecular Biology Laboratory
EST	expressed sequenced tag
Et-Br	ethidium bromid
GDP	guanosine 5'-Diphosphate
GTP	guanosine 5'-triphosphate

HCl	hydrochloric acid
HEPES	N-2-hydroxyethylpiperazine-N'-2ethanesulfonic acid
kbp	kilobase pair.
KDa	kilodalton
l	liter
M	molar
Mb	mega base pair
mg	milligram
$MgCl_2$	magnesium chloride
min	minute
ml	milliliter
mRNA	messenger ribonucleic acid.
MW	molecular weight
NaCl	sodium chloride
NaOH	sodium hydroxide
NCBI	National Center for Biotechnology Information
ORF	open reading frame
PCR	Polymerase chain reaction
PHI-BLAST	Pattern Hit Initiated BLAST (Basic Local Alignment Search Tool)
***psa*M**	M polypeptide of photosystem I
***rbc*L**	ribulose-bisphosphate carboxylase large subunit
RNA	ribonucleic acid.
RNAse I	Ribonulcease A type I-A
RNase	ribonuclease
SDS	sodium dodecyl sulfate
SMART	Simple modular architecture research tool
ssDNA	single strand deoxyribonucleic acid
ssRNA	single strand ribonucleic acid
Taq	*Thermus auquaticus*

TBS	trisbuffered saline
TE	buffer Tris and EDTA buffer
TEMED	N,N,N.,N’-tetramethyl-ethylendiamine
Tris	2-amino-2(hydroxymethel)propane-1,3-diol
tRNA	Transfer ribonucleic acid
U	unit
ycf	hypothetical chloroplast ORF

1. INTRODUCTION

1.1 The principal characteristics of the Cryptophyta cell

The Cryptophyta (crypto means "hidden" and phyta means "plant [phylum name]"), the phylum of which individuals are called cryptophytes, or cryptomonads (monad means "single object [unicell]") is a small group of flagellates, distributing in fresh, salt, and brackish water environments. This division consists of more than 20 genera (*Chroomonas, Cryptomonas, Falcomonas, Geminigera, Goniomonas, Guillardia, Hemiselmis, Plagioselmis, Proteomonas, Storeatula, Rhodomonas, Teleaulax etc…*) and approximately 200 species, half of which living in freshwater and the rest in marine environment (van den Hoek, 1996; Lee, 1999; Clay *et al.*, 2001; Hoef-Emden & Melkonian, 2003; Hoef-Emden, 2005a; Hoef-Emden, 2007, Hoef-Emden and Archibald, 2008a).

The cryptophyte cell appearances are ovoid to sole-shape, around 10 – 50 µm in size and have a dorsoventrally motif with a convex shape in the dorsal side and a flattened one in the ventral side (see Fig. 1.1.1) - ventral faces being defined by with the cell is invaginated (Hoef-Emden and Archibald, 2008a).

Cells are naked, lack of true cell wall, but covered by periplast, a sandwich-layered structure which is proteinaceous and is subdivided into inner and surface periplast components (IPC and SPC, respectively) with the plasma membrane in between. The IPC contains protein and consist of single sheet or multiple plates which have various shapes. The SPC, which is immediately outside of plasma membrane, contains plates, heptagonal rosette scales, mucilage or a combination of any of these (Hoef-Emden and Archibald, 2008a).

They have two unequal flagella emerging from above a deep furrow-gullet system (cells have a slit-like ventral opening or tubular invagination, respectively) located on the ventral side of the cell. The wall of this system is lined by numerous ejectosomes (explosive organelles) similar to trichocysts.

Some genera, such as *Cryptomonas*, have a combination of furrow and gullet, while other genera may have either only gullet or only a furrow. The flagella of cryptophytes, which differ in length, resemble those of stramenoplile. Two opposite rows of stiff flagellar hairs (mastigonomes) are carried by longer flagellum, while only single shorter row is in the shorter flagellum (Hoef-Emden and Archibald, 2008).

Most of autotrophic Cryptophyta contain one or two parietal plastids in bi-lobed and/or H-shaped form response for photosynthetic activities while phagoptrophic genus *Goniomonas* has no plastid. In the heterotrophic lineage *Cryptomonas paramaecium* [formerly *Chilomonas paramaecium* (Hoef-Emden & Melkonian, 2003)], the plastids are still present but are called as leukoplast due to the missing of photosynthetic pigments. The plastid is surrounded by four distinct membranes: two from chloroplast endoplasmic reticulum and two from the chloroplast envelope (Hoef-Emden and Archibald, 2008a).

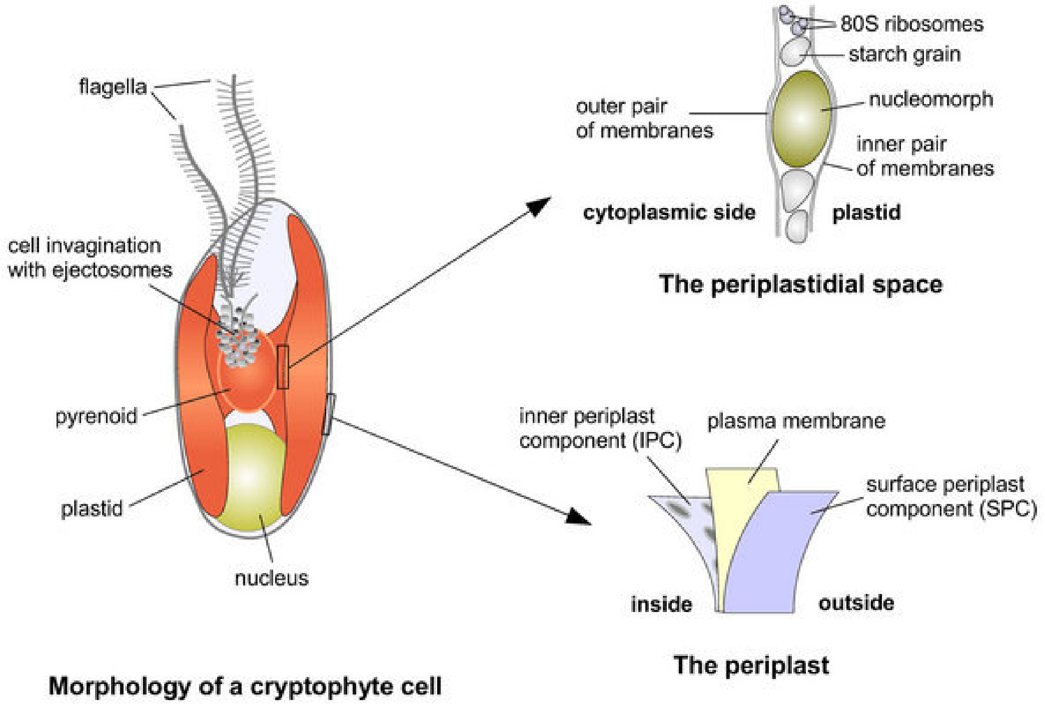
In the space between the two inner and outer membranes, called periplastidial compartment, the nucleomorph, starch granules and 80S ribosomes are located. The nucleomorph can be interpreted as the vestigial nucleus of the red algal endosymbiont that gave rise to the chloroplasts of the Cryptophyta. Inside of the plastid, the thylakoids are often arranged in pairs, to form lamella with no girdle lamellae; sometime, lamella with a larger number of thylakoids is found (Hoef-Emden and Archibald, 2008a).

A single pyrenoid (proteinaceous structure contains high amounts of paracrystalline ribulose-1,5-bisphosphate carboxylase/oxygenase) may be present centrally in the plastid, towards the dorsal side of the cell. In bi-lobed plastids this position corresponds to the bridge between the two plastid lobes. In some species, two or more pyrenoids are closely appressed to the left and right inner sides of the plastid lobes (Hoef-Emden and Archibald, 2008).

The natural colors of the cryptophytes can be blue, blue-green, reddish, red-brown, olive green, brown or yellow-brown, since they contain chlorophylls *a* and c_2 (but not chlorophyll *b*) on the outside of the thylakoids and one kind of phycobiliproteins in the thylakoids lumina, no phycobilisomes (Lee, 1999). According to Hoef-Emden (2008b), there are eight cryptophyte biliproteins named according to their absorption maximum: PE (phycoerythrin) 545, PE 555, PE 566, PC (phycocyanin) 569, PC 577, PC 612, PC 630, PC 645.

Since the plastids of the cryptophytes were originated from a red alga, it is also termed a rhodoplast to differentiate it from the chloroplasts of the green lineage and cyanelle of glaucophyte algae derived from endosymbiotic cyanobacteria (van den Hoek, 1996).

All cryptophyte algae contain ejective organelle called ejectosome or ejectisome. They occur in two sizes in the cells: the small one is located underneath the cell surface while the larger line a sub-apically and ventrally located cell invagination. When inactivated, the ejectosomes are two connected ribbons tightly coiled like springs: the central and peripheral larger spiral ribbons. When the central tube “ejects”, it pulls the spiral tube along its body. These two ribbons will be discharged when the cell exposes to environmental stimulus and the cells try to escape from the source of danger by jumping away in a zig-zag course. The formation of ejectosome seems to be involved with Golgi complex (van den Hoek, 1996; Lee, 1999; Clay *et al.*, 2001, Hoef-Emden and Archibald, 2008a).



Morphology of a cryptophyte cell

The periplastidial space

The periplast

Fig. 1.1.1: *Morphology of the cryptophytes. © 2008 Kerstin Hoef-Emden.*

1.2 The genomes of cryptophytes

Douglas *et al.* (1990) demonstrated phylogienatically that cryptophytes have originated from two different eukaryotic cells via secondary endosymbiotic process: a heterotrophic protozoon ancestor engulfed a eukaryotic photoautotrophic alga, but instead of being digested completely the "eukaryotic photoautrophic alga" was remained and reduced to a complex plastid. As a result, four phylogentically distinct genomes are presented in the cryptophytes (see Fig. 1.2.1). The host cell contributed the nucleus and mitochondrion; the engulfed alga supported its nucleus (now so termed a nucleomorph) and chloroplast genome as well as cytosol of the red alga with ribosome and starch synthesis. While the ancestor of eukaryotic photoautotrophic alga was accepted widely to be rhodophyte alga (red alga), the ancestor of the host cell component is still in the debate (Douglas *et al.*, 2001; Cavalier-Smith, 2002, Hoef-Emden and Archibald, 2008).

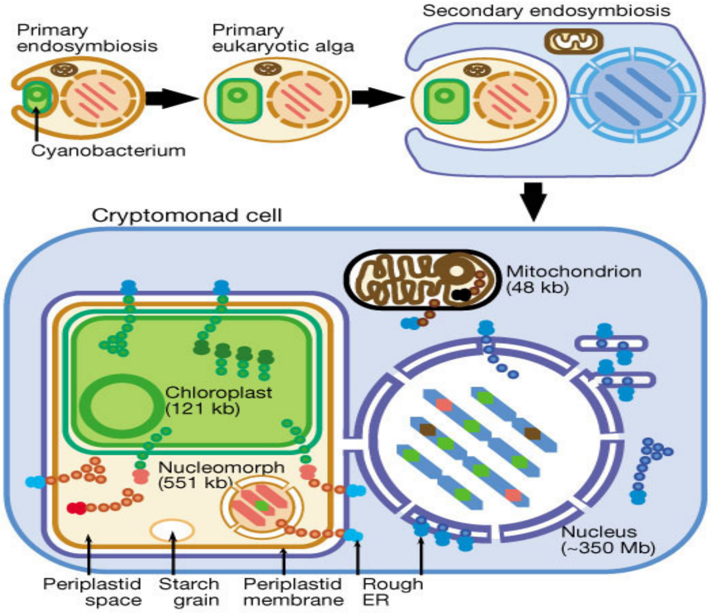


Fig. 1.2.1: *Illustration of the secondary endosymbiotic process (above) and four phylogenetically distinct genomes in the cryptophytes (copied from Douglas* et al.*, 2001).*

Genomes of chloroplast and mitochondria are double-strand circular DNA. While cryptophyte mitochondria genome sizes are around 48.5 kb and 60.5 kb [*Rhodomonas salina* (Hauth *et al.*, 2005) and *Helmiselmis andrenii* CCMP644 (Kim *et al.*, 2008), respectively], the sizes of cryptophyte plastomes are two or three time larger. Plastid genome of *G. theta* is 124.5 kb (Douglas & Penny, 1999) and *R. salina* is 135.8 kb (Khan *et al.*, 2007b) in size that are recorded the largest secondary endosymbiotic genomes thus far.

Some differences in characteristics among cryptophyte plastomes and to other plastid genomes of chromophytes and rhodophytes, mainly in the presence or absence of one or some genes, were found. For example, there are 12 genes presented in *R. salina* but not found in *G. theta*; and 8 genes are in *G. theta* but are absent from *R. salina*; HlpA gene (a histone-like protein) presented in *G. theta, R. salina, Cyanidioschyzon merolae, Galdieria sulphuraria*, but not found in haptophytes, heterokonts and nucleus-encoded in apicomlexans. The intron seems to be not an element of any chromist and red plastomes but they are still kept in *R. salina* plastome. And most surprisingly, the *R. salina* plastome acquired a gene encodes the tau/gamma components of bacterial DNA polymerase III (dnaX) by lateral gene transfer (Khan *et al.*, 2007b).

Except for genus *Goniomonas*, all cryptophytes possess small double-membrane bound organelles located in the space between chloroplast and chloroplast endoplasmic reticulum called nucleomorph**.** It has nuclear pore-like structure and an electron-dense region similar to a nucleolus. A broadly survey the karyotypic structure of nucleomorph genomes by using pulsed-field gel electrophoresis done by Lane *et al.* (2006b), Tanifuji *et al.* (2006) and Phipps *et al.* (2008) showed that the genome sizes of cryptophyte nucleomorphs are quite different. The smallest and the largest ones are 450 kb and 845 kb in colorless *C. paramaecium* CCAP 977/2a and pigmented *Rhodomonas* sp. CCMP 1178, respectively. Even though the sizes are distinct, all Cryptophyta nucleomorph always contain three small linear chromosomes. Interestingly, this feature is shared universally with an unrelated algal group, the chlorarachniophytes. Still now, there are only two fully sequenced nucleomorph genomes, one of *G. theta* and one of *Hemiselmis andersenii*, with 551 and 572 kb in size, respectively (Douglas *et al.*, 2001; Lane & Archibald, 2006a). Both have very high level of gene compaction, 1 gene/kb. They eliminated most metabolic functional genes, but kept genes encoding protein required for basic eukaryotic cellular process and 30 genes that their gene products will be transferred to chloroplast. To express the genes, cryptophyte nucleomorph genomes need hundreds of genetic-housekeeping genes. Interestingly, the sizes of proteins produced by cryptophyte nucleomorph are smaller than those in their free-living alga ancestor are. Even, in the smaller genome, *G. theta*, the producing of protein is less than that of *H. andresenii*. The evolutionary processes seem to control strongly the capacity of the nucleomorph genome in which the genome compaction occurred not only in noncoding DNA but also in coding regions. The evolution of their protein structures and functions, therefore, has been examined (Lane & Archibald, 2006a).

The host cell nucleus, lying in the posterior half of the cell, is quite large, 350 Mb in size, with 40 to 210 linear chromosomes to be counted at metaphase (van den Hoek, 1996; Lee, 1999; Douglas *et al.*, 2001). However, little molecular genetic information of the cryptophyte host nuclear genomes is available. Fortunately, Khan *et al.*, (2007a) published the partial results of the sequencing projects for two distantly related cryptomonad nuclear genomes, photoautotrophic *R. salina* CCMP 1319 and heterotrophic *C. paramaecium* CCAP 977/2a. This publication provided the first insight of the partial structure and composition of the cryptomonad nuclear genomes.

1.3 Sequencing strategies for Cryptophyta genomes

Since the first plastid physical map of *G. theta* (mislabeled as *Cryptomonas* Φ) published (Douglas, 1988) the genome studies of phylum Cryptophyta have developed strongly with the assistance of the cost-effective, high-throughput sequencing facilities.

At the first stage of genomic era, Douglas & Penny (1999) used the most common strategy to read the whole plastid genome of *G. theta*. They isolated the chloroplast DNA from total DNA by ultracentrifugation with Hoechst 33258-cesium chloride. According to density gradient, each genome distributed as independently separated bands. The plastid DNA was fractionated by upward displacement. Clear chloroplast DNA, then, was cut into small fragments by several different restriction endonucleases and cloned into suitable vectors. To get the physical map, the authors used several significantly probes containing reliable makers such as 23S rRNA, 16S rRNA, *psa*A, *rbc*L, *etc* … These clones were read for nucleotide compositions.

The advantages of genome DNA isolation by ultracentrifugation with Hoechst 33258-cesium chloride are allowing the researchers to obtain fractions of other genomes (mitochondria, nucleormorph and host nucleus) in addition to plastid genome. Hoechst 33258 or bisbenzamide is a member of a family of fluorescent stains for labeling DNA in fluorescence microscopy and fluorescent-activated cell sorting, excited by ultraviolet light at around 350 nm. This dye associates with AT-rich portions of the DNA and causes regions of DNA with higher proportions of AT to be buoyed up in the gradient (Sambrook & Russell, 2006). Utilizing this, Douglas *et al.* (2001) published the fully nucleotide compositions of *G. theta* nucleomorph genome with DNA material obtained from the above centrifugation process.

However, one of the challenges of this approach is the contamination. Because the bands of genome DNA are very close together, the contamination among genomes cannot be avoided when pick up a certain genome DNA (Jansen *et al.*, 2005).

To ovoid the problems of contamination, the shotgun strategy seems to be more powerful in sequencing projects (Sterky & Lundeberg, 2000; Jansen *et al.*, 2005). In this approach, the ultracentrifugation is ignored but total DNA can get by any standard methods, and then sheared them into fragments with the size up to 150 kb. The 40 – 50 kb or/and 100 – 150 kb are separated by pulse field gel electrophoresis, and then collect for cloning into Fosmid or BAC, respectively. Using the hybridization techniques, in which the probes prepared by PCR technique of the marker genes distributed throughout the plastome, the chloroplast-gene-holding fragments can be recognized. A minimal collection of Formids that covers the whole plastid genome is obtained and then the inserts are read by any available sequencing techniques. In case of using BAC library, very few of clones are needed to read the whole plastome, dependence on its size. This strategy was employed by Khan *et al.*, (2007b) and Lane & Archibald (2006a) for sequencing the entire *R. salina* CCMP 1319 plastome and *H. andersenii* nucleomorph genome. The host nuclear genome sequencing projects of photosynthetic *R. salina* CCMP 1319 and heterotrophic *C. paramaecium* CCAP 977/2a are now ongoing but the partial results using shotgun approach were published (Khan *et al.*, 2007a).

One of the advantages of the shotgun sequencing is that all genomes of certain examined organism can be cloned into Fosmid or/and BAC libraries. The clones, then, are stored for further investigation. Secondly, the readings based on the well-chosen Fosmid or BAC libraries help to save time at later stages of sequencing projects such as base-calling and assembling the inserts (Jansen *et al.*, 2005).

Generally, the plastome sizes are extremely small in comparison with total cellular DNA, therefore the ratio of nonplastid-to-plastid clones are very high. Thus, a great number of clones are needed for screening to guarantee enough plastid clones will be get to cover the whole plastome. Besides, the shotgun strategy requires more time, cost, labors and high-tech instruments (Jansen *et al.*, 2005).

Even though the above strategies helped the researchers harvest a huge amount of molecular information from interesting Cryptophyta genomes, they still have had some limitations as pointed out. Specially, both approaches require the remarkable initial materials. It is impossible for rare or hard to culture subjects, for example colorless *Cryptomonas* SAG 977-2f.

Another approach for collecting genome structures and sequences is the combination of long-range PCR and primer-walking strategies. At first, a set of large fragments (up to 20 kb) covering the whole plastomes are amplified by using highly conserved chloroplast primers. The amplicons, then, are used to produce smaller segments for cloning or reading directly by any sequencing methods. The first DNA reading is started with two universal primers at two ends of the segments. The newly obtained DNA databases are used to construct new sequencing primers for the second reaction. This process is continued until the whole nucleotide compositions of interesting fragments are fully read on both strands. In this approach, less initial DNA samples are needed (Ponce & Micol, 1992; Cheng *et al.*, 1994; Jansen *et al.*, 2005).

The combination of long-range PCR and primer-walking approaches was mainly applied in mitochondria sequencing projects, especially in animal mitochondria since the mitochondrial sizes were about 16 kb in length. The researchers using this strategy valued that it helped to collect entire mitochondrial sequences more rapidly than traditional methods, e.g. cloning the mitochondrial DNA into suitable vectors and sequencing them. Moreover, this approach were very useful in case the total DNA was not enough (up to hundreds of milligram) to subject for ultracentrifugation as in traditional strategies. In addition, this approach required standard molecular biology facilities only (Giesecke *et al.*, 1992; Hu *et al.*, 2002; Yamauchi *et al.*, 2004a;

Yamauchi *et al.*, 2004b; Kim *et al.*, 2005; Munemasa *et al.*, 2006; Mereu *et al.*, 2007).

However, the long-range PCR combined with primer-walking sequencing approach seemed to be un-favorite to chloroplast genomic researchers. Almost only one group of researchers employed long-range PCR to sequence three basal angiosperm genomes (Goremykin *et al.*, 2003a; Goremykin *et al.*, 2003b; Goremykin *et al.*, 2004). Geromykin's group isolated the entire chloroplast plastome by long-range PCR techniques. The products with the size up to 20 kb, then, were sheared by nebulization to get the smaller fragments of 0.5 to 1.5 kb in length, which were sub-cloned to suitable vectors for reading nucleotide compositions. This strategy seems to be a modified version of shotgun approach, in which the total DNA isolation step is replaced by long-range PCR to cover plastid genome only.

Some disadvantages of this approach were pointed out by Jansen *et al.* (2005)

> (1) The primers for long-range PCR may be useless if plastomes changed their gene orders or the primers cannot hit expected sites due to the substantial sequences divergence;
>
> (2) This approach replies on PCR technique which can sometime give negative results or false positive in certain DNA fragments;
>
> (3) The designing primers for long-range PCR could be problematic if the chloroplast genome information of more or less relative taxa are not available;
>
> (4) Time-consuming may be a minus when comparison with other available methods.

Investigating the completely sequenced chloroplast genomes from Cyanobacteria to higher plant showed that there is core set of 45 genes retained in all taxa (Martin *et al.*, 1998; Martin *et al.*, 2002; Grzebyk *et al.*, 2003; Nozaki *et al.*, 2003). This gene list is extremely useful when applied to sequence certain chloroplast(s). The construction of universal primers for chloroplast is thankfully easier. Subsequently, the primers can be combined together in the most optimal manners to increase the probability of amplifying the fragment in between the two selected genes. The amplicons, afterward, can be sequenced following the available standard protocols. Hence, the disadvantages of the long-range PCR combined with primer-walking sequencing approach seem to be overcome.

1.4 Aims of study

Traditionally, the common approach in plastid phylogenetic studies is that sequence one or several highly conserved genes such as ribosomal operon, transfer RNAs and protein – coding genes, then construct the phylogenies based on single or concatenated data sets (Hoef-Emden, 2005b). Another approach is that using whole plastid genomes to analyze the molecular evolution (Martin *et al.*, 1998; Martin *et al.*, 2002; Grzebyk *et al.*, 2003; Nozaki *et al.*, 2003). The phylogenetic relationship among cryptophyte algae was well established by applying the former approach (Marin *et al.*, 1998; Clay *et al.*, 1999; Deane *et al.*, 2002; Hoef-Emden *et al.*, 2002; Hoef-Emden & Melkonian, 2003; Hoef-Emden, 2005a; Hoef-Emden, 2007).

Hoef-Emden (2005a) used both the partial nuclear and nucleomorph ribosomal operon phylogenies to show a new and unexpected finding that at least three different colorless lineages were found in the genus *Cryptomonas*. This raised questions about the evolutionary history of these interesting objects and their relatives as well as the role of plastid genomes such as whether these three lineages resulted from similar or from different evolutionary events or what are the mutual relationship or/and roles of photosynthetic genes in the absence of photosynthetic activities, *etc*.

In the first stage of the thesis, the *rbc*L genes (1,5-biphosphate carboxylase/oxygenase [RuBisCO] large subunit) was chosen for searching and sequencing among photosynthetic and heterotrophic *Cryptomonas* species. The data set was then done a phylogenetic analysis to compare with those of nuclear rDNA (concatenated SSU rDNA, ITS2 and partial LSU rDNA) and nucleomorph SSU rDNA. The *rbc*L sequences also were subjected to examine the codon usage of two-fold degenerate NNY codons to deduce the differences of functional constraints and expression levels in this gene in the genus *Cryptomonas* (the later works were done by Dr. Kerstin Hoef-Emden).

In the next stage, an ambition for investing deeply the evolutionary pathways of the colorless plastids by comparison of their plastome information to those of pigmented relatives was planned. In the proposal, strains *C. paramaecium* 977/2a and *C. oobovoidea* CCAC 0031 (representative for a heterotrophic and autotrophic *Cryptomonas,* respectively) were chosen to isolate by ultracentrifugation with Hoechst 33258-cesium chloride then sequence by shotgun approach. However, the plastid genome isolation was unsuccessful repeatedly due to the technical problems. To save the time, the object has been changed in which the fragment from 16 rRNA to *rbc*L genes (around 7 kb) was replaced for whole plastid genomes. Another significant alternation was that long-range PCR and primer-walking approaches were applied to read these fragments. The numbers of taxon sampling also increased to 22 strains in which five strains were not *Cryptomonas*.

The protein-coding genes harvested from these fragments were used as alternative makers for *Cryptomonas* phylogenies to compare with other previously results published recently. The study also attempts to find new more evidence to explain the relationship between the changing from autotrophic to heterotrophic lifestyle in colorless *Cryptomonas* lineages.

2. MATERIALS AND METHODS

2.1 Chemicals and medium

β-mercaptoethanol	Sigma
Ammonium acetate	Merck
Bromophenol blue	Merck
Agarose	Invitrogen
Amoniumpersulfat (APS)	Sigma
Biotin	Serva
$Ca(NO_3)_2.4H_2O$	Merck
Chloroform	Applichem
$CoCl_2.6H_2O$	Merck
CTAB (Cetyltrimethyl ammonium bromide)	Fluka
DMSO	Merck or/and Roth
EDTA (Titriplex III)	Serva
Ethanol (pure for analysis 99.8%)	Roth
Ethidiumbromid 1%	Applichem
$FeSO_4.7H_2O$	Merck
Glacial acetic acid	Merck
Glycerol	Sigma
H_3BO_3	Merck
HCl	Merck
HEPES	Roth
Isoamylalcohol	Merck
KNO_3	Merck
KOH	Merck

Lab-Lemco Broth	Oxoid
$MgSO_4.7H_2O$	Merck
$MnCl_2.4H_2O$	Merck
NaCl	Merck
Niacinamide	Sigma
Sodium acetate trihydrate	Merck
$(NH_4)_2HPO_4$	Merck
TEMED (N,N,N’, N’-Tetramethylethylendiamin)	Fluka
Thiamine-HCl	Serva
Tris,	Invitrogen
Vitamin B_{12}	Serva
Urea	Sigma
Xylene cyanol	Merck
$ZnSO_4.7H_2O$	Merck

2.2 Algal culture medium

Thirty-one cryptophyte strains used in this study were listed in the Table 2.1. Their participating in the first part or/and second part of the project (*rbc*L and 16S rRNA-*rbc*L fragments, respectively) were indicated clearly in the Table 2.2.1.

Five colorless strains were grown in CHM medium at 15^{o}C without light. The photoautotrophic strains were grown in WARIS-H medium (Kies, 1967: modified, (McFadden & Melkonian, 1986) at 15^{o}C, under 70 µmol photon x m^2 x sec^{-1} light with a 14:10h light: dark cycle. **S**tocks were checked every week to be sure contamination free, when the cells reached high density, they were transfer into sterile Erlenmeyer flasks containing 75 ml new autoclaved medium.

WARIS-H medium

Stock solutions and final concentration in culture medium	***Addition per 1 litre stock solution***	***Addition per 1 litre culture medium***
1. KNO_3 (1.00 mM)	100.00 g	1 ml
2. $MgSO_4.7H_2O$ (81.10 µM)	20.00 g	1 ml
3. $(NH_4)_2HPO_4$ (0.15 mM)	20.00 g	1 ml
4. $Ca(NO_3)_2.4H_2O$ (0.42 mM)	100.00 g	1 ml
5. HEPES (1.00 mM)	238.31 g	1 ml
6. P-II Metals		1 ml
EDTA (Titriplex III) (8.06 µM)	3.00 g	
H_3BO_3 (18.43 µM)	1.14 g	
$MnCl_2$ .4 H_2O (0.73 µM)	144.00 mg	
$ZnSO_4$. 7 H_2O (73.00 nM)	21.00 mg	
$CoCl_2$. 6 H_2O (16.80 nM)	4.00 mg	
Dissolve EDTA (Titriplex III) and boric acid in bidistilled H_2O, then add metals one after the other.		
7. Fe-EDTA		1 ml
EDTA (Titriplex II)(17.86 µM)	5.22 g	
$FeSO_4.7H_2O$ (17.90 µM)	4.98 g	
1 N KOH	54.00 ml	
EDTA (Titriplex II) and $FeSO_4.7H_2O$ is heated for 30 min (100°C); KOH is added to		

the cooled mixture

8. Vitamins		1 ml
Vitamin B_{12} (0.15 nM)	0.20 mg	
Biotin (4.10 nM)	1.00 mg	
Thiamine-HCl (0.30 µM)	100.00 mg	
Niacinamide (0.80 nM)	0.10 mg	
pH of this solution should be around pH 7.0		
9. Soil Extrac		10 ml

Preparations of stock solutions were done by technical in the lab.

Preparation of soil extract: 10g of garden-soil was mixed with 120 ml Super Q water and boiled for 10 minutes. Afterwards it was centrifuged for 10 minutes (low speed), and the supernatant was filtered through a series of membrane filters from 1.2 µm – 0.1 µm pore size. The remaining filtrate was adjusted to 100 ml with bi-distilled water. Aliquots of 10 ml were stored frozen. The soil should not be recently fertilized and should not contain too much humus.

Preparation of culture solution: Added 1 ml of stock solutions 1-8 to 1000 ml of bidistilled water. Added 1 ml of thawed soil extract (stock solution 9); adjusted the pH to 7.0 and autoclaved.

CHM (Chilomonas medium)

Sodium acetate trihydrate	1g
Lab-Lemco Broth	1g
Water	1 l

Autoclave and keep in cold room (15 °C) at least 3 days before use.

Table 2.2.1: List of Cryptophyta strains were used in this study

	Strain	***rbc*L project**	**16S rRNA-*rbc*L fragment project**
01	*C. commutata* CCAC 0109 (M0739)	Y	Y
02	*C. erosa* CCAC 0018 (M0788)	N	Y
03	*C. loricata* M2088	N	Y
04	*C. oobovoidea* CCAC 0031 (M1094)	Y	Y
05	*C. oobovoidea* CCAP 979/46	Y	N
06	*C. curvata* CCAC 0006 (M0420)	N	Y
07	*C. borealis* CCAC 0113 (M1083)	Y	Y
08	*C. borealis* SCCAP K-0063	Y	N
09	*C. gyropyrenoidosa* sp. nov. CCAC 0108 (M1079)	Y	Y
10	*C. lundii* CCAC 0107 (M0850)	N	Y
11	*C. marssonii* CCAC 0086	Y	Y
12	C. *marssonii* CCAC 0103 (M1475)	Y	N
13	*C. ovata* CCAC 0064 (M0847)	Y	Y
14	*C. paramaecium* CCAP 977/1	Y	N
15	*C. paramaecium* CCAP 977/2a	N	Y
16	*C. paramaecium* CCAC 0056 (M1303)	Y	Y
17	*C. paramaecium* M2452	Y	Y
18	*C. paramaecium* M2180	Y	N

19	*C.* sp. M1634	Y	Y
20	*C.* SAG 977-2f	Y	Y
21	*C. phaseolus* SAG 2013	N	Y
22	*C. pyrenoidifera* CCAP 977/61	Y	N
23	*C. pyrenoidifera* CCMP 0152	Y	Y
24	*C. pyrenoidifera* M1077	Y	N
25	*C. tetrapyrenoidosa* M1092	Y	Y
26	*C. tetrapyrenoidosa* NIES 279	Y	N
27	*Chroomonas* sp. SAG 980-1	N	Y
28	*Hemiselmis tepida* CCMP 0443	N	Y
29	*Proteomonas* sp. CCPM 0704	N	Y
30	*Rhodomonas* sp. M1480	N	Y
31	*Teleaulax* sp. SCCAP K-416	N	Y

Note: Y – yes, participated in the project; N – No participated in the project

CCAP: Culture Collection of Algae and Protozoa (UK); M: Algae Culture Collection Melkonian at the University of Cologne (Germany); CCAC Culture Collection of Algae at the University of Cologne (Germany), SCCAP: Scandinavian Culture Centre for Algae and Protozoa at the University of Copenhagen (Denmark); NIES: Microbial Culture Collection at the National Institute for Environmental Studies (Japan). SAG (Sammlung von Algenkulturen der Universität Göttingen, Germany), species names according to Hoef-emden & Melkonian (2003), Hoef-Emden (2007) and Land & Archibald (2008).

2.3 Harvest the cells for DNA isolation

In general, cells were harvested at the high density (10^5 - 10^6 cells/ml) empirically. Two ml of stock culture were poured into one sterile Eppendorf tube then centrifuged at 1000 *g* for 10-20 minutes (Beckman Coulter, Krefeld, Germany). The pellet was washed with fresh culture medium or dH_2O to remove old culture medium, centrifuged again. The supernatant was removed completely while the pellet was stored by liquid nitrogen or in deep cold (-70°C) at least 30 minutes before continuing the isolation of total DNA.

2.4 Isolation of total genomic DNA by CTAB method

To quickly isolate total DNA, the pellet obtained above was processed either by DNAEasy Plant MiniKit (Qiagen) according to manufacturer's procedure or CTAB method (Hoef-Emden *et al.*, 2002 modified from Doyle & Doyle, 1990). Suspended frozen cells in 2X CTAB-buffer and 2 µl β-mercaptoethanol. Vortexed immediately and incubated in 60 °C (water-bath), for 10 – 15 minutes (β-mercaptoethanol broke the S-S bonds while CTAB formed a soluble complex with the DNA in the presence of high salt; cell wall debris, denatured proteins, and polysaccharides were removed by extracting the aqueous phase with chloroform). Added 1 ml chloroform-isoamylalcohol, then wrapped tube with aluminum, shook (inverted) gently the tube for 10 minutes continuously at room temperature. Centrifuged by small table centrifuger for 2 – 3 minutes at maximum speed in order to separate the sample into two phases (organic/ aqueous). Transferred the aqueous phase (containing the DNA) into new-labeled 2ml- Eppendorf tube, without traces of the other phases (pipette slowly helps with this); incubated on ice; filled up volume with water to 1 ml. Repeated these steps 3-5 times to remove completely contaminants. Added 0.7 ml Isopropanol to precipitate DNA, immediately inverted tube several times, incubated at -20°C for 10 minutes (optimally, DNA-flakes should be visible). Centrifuged the tube at 15300 rpm for 5 – 10 minutes by cooled desk-centrifuge at 4°C (Sigma 2K15), discarded supernatant very carefully (colorless pellet was usually visible). Added 1 ml of cold 70 – 80% Ethanol (- 20°C), inverted gently, then incubated for 5 min (-20°C). Centrifuged again, discarded the supernatant; sometimes, the DNA-pellet was no longer firmly attached, and swam somewhere. Opened the tube and dried the pellet at a clean place at room temperature, 30 minutes to 2 h, pellet had to be dried (no traces of ethanol).

Added 1 ml TE-buffer to dilute total genomic DNA (working on ice 4°C), transferred to new free-DNA Eppendorf tube (if necessary; mixed for some time). Applied RNase-I (Ribonulcease A type I-A from bovine pancreas, Sigma-Aldrich, R4875) to destroy unnecessary RNA (final concentration should be 10 µg/ml), then incubated for 30 minutes to 1 h at 37°C, mixed gently during the reaction. Transferred the solution to a new free-DNA 5ml Falcon tube before added 2 ml ammonium acetate 2.5 M and 2.5 ml **c**old (4°C) absolute ethanol; mixed well and incubated again for 1 – 2h or overnight at -20 °C. Longer time would tend to yield more DNA, but also more contaminants. Centrifuged at 10000 *g* for 10 minutes, 4 °C. Discarded carefully the supernatant and washed by 80% ethanol (cold) for 10 minutes; transferred to new sterile 2 ml tube; centrifuged again for 5 minutes at 10000 *g* 4 °C. Discarded carefully the supernatant; dried 30 – 60 minutes at the clean air. Added 1 ml TE buffer, mixed by pipette gently then stored at - 20 °C until use.

The DNA concentration was determinate by monitoring absorbance at 260 and 280 nm in a UV spectrophotometer (Eppendorf).

CTAB buffer for 1l:

100 ml	of 1 M Tris, pH 8.0
280 ml	of 5 M NaCl
40 ml	of 0.5 M EDTA
20 g	of CTAB (Cetyltrimethyl ammonium bromide)

Ethanol:

Absolute

Cold 90%, 80% and 70% (store at - 20 °C)

CI Chloroform-Isoamylalcohol (24:1 v/v)

Chloroform	24 ml
Isoamylalcohol	1 ml

store at 4°C in dark glass bottle.

Ammoniumacetate 2.5 M:

250 ml of Ammoniumacetate 10 M

TE buffer for 1l use:

10 ml of 1 M Tris, pH 8.0

2 ml of 0.5 M EDTA

1 M Tris, pH 8.0 for 1 l:

121.1 g Tris

700 ml ddH_2O dissolve Tris and bring to 900 ml.

pH to 8.0 with concentrated HCl (will need ~50ml) and bring to 1 l.

0.5 M EDTA pH 8.0 for 1l:

186.12 g of EDTA

750 ml ddH_2O add about 20 g of NaOH pellets

slowly add more NaOH until pH is 8.0,

EDTA will not dissolve until the pH is near 8.0.

5 M NaCl for 1l

292.2 g of NaCl

700 ml ddH_2O dissolve and bring to 1 L.

Ammoniumacetate 10 M:

Ammonium acetate	770 g
Glacial acetic acid	800 ml

Add distilled H_2O to make a final volume of 1 liter.

2.5 Amplify the plastid *rbc*L gene of *Cryptomonas* strains by BioTherm™ Taq DNA Polymerase and read their DNA compositions by SequiTherm EXCEL™ II DNA Sequencing Kit-LC and Li-Cor 4200L bidirectional sequencer

At the first stage of the project, the Cryptophyta plastid *rbc*L gene was chosen to read its nucleotide composition prior phylogenetic analyses.

2.5.1 Construct primers

The rhodophyte/cryptophyte-specific *rbc*L primers were constructed manually by comparing and looking for highly conserved domains of around 100 *rbc*L genes of cryptophytes, rhodophytes and chlorophytes obtained from the EMBL/GeneBank database and were purchased from Metabion (Germany). The PCR products and length of sequenced products were expected around 1.3 to 1.4 kb. The names and approximate positions of those primers were given in Table 2.5.1.1 and Fig. 2.5.1.1.

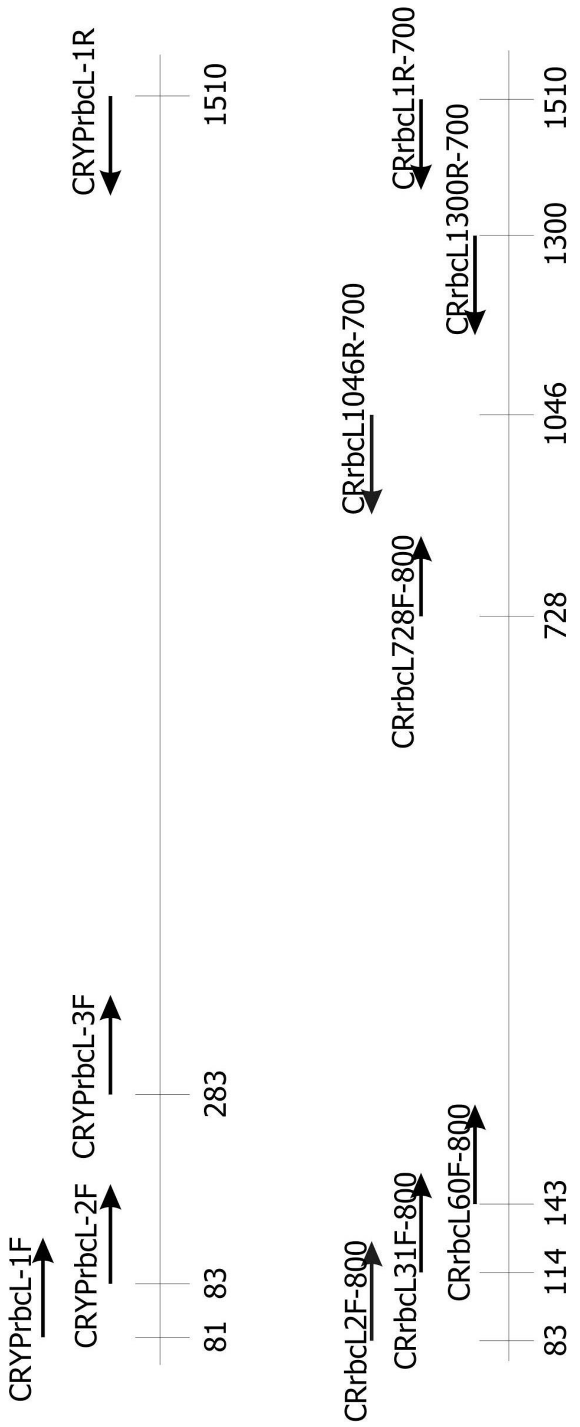


Fig. 2.5.1.1: *Approximate positions of the PCR (above) and sequencing primers (below). The positions had been aligned following the sequence of* G. theta *(AF041468).*

Table 2.5.1.1: PCR and sequencing primers for amplifying and sequencing cryptophyte *rbc*L gene

PCR PRIMERS	
Name	**SEQUENCE (5' to 3')**
CRYP*rbc*L1F	CAA GGA GGA AWA YAT GTC TCA ATC
CRYP*rbc*L2F	AGG AGG AAW AYA TGT CTC CTC AAT CCG
CRYP*rbc*L3F	GAA TCT TCA ACA GCA ACW TGG AC
CRYP*rbc*L1Rbiot	5'biotin-TCA GCT GTA TCW GTA GAA GC
SEQUENCING PRIMERS	
CR*rbc*L1R-700	IRD700-TCA GCT GTA TCW GTA GAA GC
CR*rbc*L1046R-700	IRD700-ACC WGC CAT RCG CAT CCA CTT AC
CR*rbc*L1300R-700	IRD700-TCT ARA GCY GTY ZGA AGA GGW CCA
CR*rbc*L2F-800	IRD800-AGG AGG AAW AYA TGT CTC AAT CCG
CR*rbc*L31F-800	IRD800-CTA AAT CCG TWG AAW CRC GGA CTCG
CR*rbc*L60F-800	IRD800-AAC GAA CGT TAT GAA TCA GGY G
CR*rbc*L728-800	IRD800-CTC CAR CCW TTY ATG AGA TGG

Note: CRYPsL1Rbiot was labelled at its 5'-terminus with biotin and all sequencing primers were labelled with the 700 nm (reverse primers) or 800 nm (forward primer) IR-fluorescent dye.

2.5.2 Set up conditions for PCR running

Component for each PCR reaction:

Reverse primer (10 pmol/µl):	0,25
Forward primer (10 pmol/µl):	0,25
dNTP (2mM) :	2,5
Buffer 10X:	2,5
Water:	17,5
BioTherm™ Taq DNA Polymerase (5UI/µl):	0,5
DNA template (50-100 ng/µl)	1.5

Reactions were run in an MWG biotech thermal cycler following by THD-PCR3 or THD-PCR5 program:

THD-PCR3

Initially denature the template: 95^{o}C for 3’
Followed 30 cycles:
    Denature at 95 ^{o}C for 2’
    Anneal at 45 ^{o}C for 2’
    Extend at 68 ^{o}C for 3’
Ended with 68 ^{o}C for 3’

THD-PCR4

Initially denature the template: 96^{o}C for 3’
Followed 30 cycles:
    Denature at 95 ^{o}C for 1’
    Anneal at 52 ^{o}C for 2’
    Extend at 68 ^{o}C for 3’
Ended with 68 ^{o}C for 5’
After amplification, the samples might be kept at 4 ^{o}C overnight.

2.5.3 Purify PCR products by use streptavidin-coated dynabeads M-280 system

After the first PCR reaction using biotinylated primers, the amplified biotinylated DNA fragments were isolated by apply Dynabead M-280 system (Dynal, Olso, Norway). Took 5 µl of streptavidin-coated dynabeads solution for 1 PCR product. Put on magnetic field for 1-2 minutes, removed supernatant completely, and did not touch precipitate. Added 5 µl washing buffer, mixed gently and put on magnetic field again, discarded supernatant as before. Repeated washing step several times. Took PCR products (at least 5 µl) and mixed with washed streptavidin-coated dynabeads. Mixed well and laid on the table for 45 minutes for biotin-streptavidin reaction to take place. Mixed several times during this process to avoid partial reaction between PCR products and streptavidin-coated dynabeads.

2.5.4 Prepare Pre-Mix and Mix for PCR-based sequencing

Labelled 200 µl-PCR tubes with C, A, T, G for ddCTP, ddATP, ddTTP, ddGTP, respectively. Put 1 µl of each dNTP into correspond tube.

Prepared Pre-Mix for each reaction:

Water:	3,675 µl;
SequiTherm EXCEL II Sequencing Buffer:	3,825 µl;
SequiTherm EXCEL II DNA Polymerase (5 U/µl):	0.5 µl;
Primer Forward or Primer reverse(1.0 pmol/ µl):	1 µl;
Total volume:	9 µl.

Mixed well, put on ice and in dark light condition until use.

See Table 2.5.1.1 for the detail the names, compositions and binding sites of sequencing primers.

After reaction between PCR products and streptavidin-coated dynabeads had been completed, the precipitated complex was washed as before with washing-buffer 3-5 times. The washing-buffer was removed by adding 5 µl of denionized water to the last precipitate then mixed well and poured out water completely. Pre-Mix was added immediately poured and pipette carefully. The last solution was Mix; it was already for use. One point eight µl of Mix was then distributed into each of labelled PCR tube containing 1 µl of dNTP. These tubes were laid immediately at 5 ^{o}C before applied to PCR machine.

Program the thermal cycler for sequencing

KHE-SEQ1

Initially denature the template: 95 ^{o}C for 2’
Followed 30 cycles:
  Denature at 94 ^{o}C for 30”
  Anneal at 40 ^{o}C for 30”
  Extend at 70 ^{o}C for 1’
Ended with 69 ^{o}C for 5’

Reactions were run in an MWG biotech thermal cycler, after completion of PCR-sequencing, applied 1.5 µl of red buffer into each PCR tube to kill enzyme. Stored at -20 ^{o}C before reading sequences.

2.5.5 Read the nucleotide compositions

Double-stranded sequences were determined with Sanger sequencing techniques using SequiTherm EXCEL™ II DNA Sequencing Kit-LC for 66 cm gels according to manufacturer’s protocol (Epicenter Technologies) and a Li-Cor 4200L bidirectional sequencer (Li-Cor Biosciences, Bad Homburg, Germany).

Electrophoresis conditions were set up Voltage (V): 2000; Current (mA): 35; Power (W): 50; Temperature (^{o}C): 45; Time remaining: 30 minutes. The gel had been pre-running for 30 minutes, while the samples were prepared in denaturation step by putting them into heat lock at 80 ^{o}C for 10-15 seconds. When denaturation step had finished, the red buffer was applied for the sample (1µ/tube). Run progress was taken place for 18 hours.

2.5.6 Proofread newly sequenced **rbc***L gene*

The AlignIR 2.0 (Licor-Biotechnology, Germany) program was used to assemble and proofread the sequences obtained. The multi sequence alignment editor SeaView (Galtier *et al.*, 1996) was applied for the manual alignment of *rbc*L gene.

2.5.7 Deposit the sequence data

The newly Cryptophyta plastid *rbc*L sequences were submitted to gene bank to obtain the accession numbers by Dr. Kerstin Hoef-Emden.

2.6 Amplify the the cryptophyte plastome 16S rRNA-*rbc*L fragments by MasterAmpTM Extra-long PCR kit and read their DNA sequences by BigDye Terminator v1.1 Cycle sequencing kit and automated ABI3730 sequencer

In the second stage of the project, the long fragments in between 16S rRNA and *rbc*L genes (around 7 kb) of 22 cryptophytes were sequenced for some further phylogenetic analyses. At first, the long fragments were amplified by using MasterAmp^{TM} Extra-long PCR kit. The long PCR products were then re-amplified by BioTherm™ Taq DNA Polymerase to produce the two short overlapping fragments (around 4 – 4.5 kb/fragment each) that would be used as the templates for reading the DNA sequences by automated ABI3730 system.

The 16S rRNA–*rbc*L fragment was chosen as potential object because it satisfied following criterions:

- Having some essential gene categories according to their functions: There are 4 gene families observed in the 16S rRNA–*rbc*L fragments of *G. theta* and *R. salina* plastome: genes family for translational machinery (16S rRNA and *rps*4 gene); gene coding for photosynthetic apparatus (*psa*M); genes participate in biosynthetic process (*chl*I and *rbc*L); hypothetical or miscellaneous gene family (*ycf*26, ORF403); and tRNA genes. Having some essential gene families allows later analyzing make more sense.

- Allowing to design universal primers for long-range PCR conveniently: Database of nucleotide sequences 23S rRNA, 16S rRNA and *rbc*L gene for Cryptophyta are now available in many gene banks (Hoef-Emden *et al.*, 2002; Hoef-Emden & Melkonian, 2003; Hoef-Emden *et al.*, 2005; Hoef-Emden, 2005a). They are excellent resources for designing the primers for long-PCR.

2.6.1 Design the primers for long-PCR

To amplify the long fragment, primers for long PCR about 25 – 30 bp long were requested. They were constructed according to some standard considerations (Palumbi, 1990):

- Hit to highly conserved domains of ribosomal RNA, transfer RNA and protein-coding genes;
- Contain some G/C at their 3' terminus to enhance the primer – template annealing step at those positions.

To design the universal long primers binding at upstream end of *rbc*L gene, the *rbc*L gene database of genus *Cryptomonas* harvested from the first part of the project was utilized; the primers were named CRYP*rbc*L2R and CRYP*rbc*L3R, accordingly.

To construct the conserved primer starting from 23S rRNA gene, the initial sequence database of cryptophyte ribosomal gene was needed. To fulfill this requirement, five *Cryptomonas* strains – CCACP 977/2a, M1634 (colorless strain), CCAC 0031, CCAC 0109 and CCAC 0006 – were randomly chosen to sequence a small section of plastid 23S rRNA gene. These obtained sequences were imported into Dr. Birger Marin's plastid and Cyanobacterial ribosomal operon database for aligning. Two primers for cryptophyte plastid rRNA operon were designed based on the highly conversed regions: domain C-D and helix G20 in 23S rRNA. They were named pt1Rlong and G20Rlong, respectively.

Approximately positions, lengths and synthetic directions of long-primers were listed and illustrated in Table 2.6.1.1 and Fig. 2.6.1.1.

Table 2.6.1.1: Primers for long-range PCR to amplify the cryptophyte plastid 23S rRNA-*rbc*L fragments

pt1Rlong:

GCCACTRCCTAYAAGTCGCCGGCTCATTCTTCAAC

from 585th to 620th position of *G. theta* platid 23S rRNA gene

G20Rlong:

CTCTARCGCCTRCACCGGATATGGACCGAACTGTC

from 2574th to 2609th position of *G. theta* platid 23S rRNA gene

CRYPrbcL2R:

ATCWGTAGAAGCRTARTTRAAHGTDATRTCTTTCC

end of *rbc*L; overlapping with CRYP*rbc*L1Rbiot (Hoef-Emden *et al.*, 2005)

CRYPrbcL3R:

GCTTGRATACCRTCWGGRTGWCCWATWGTACCMCCACC

about 100 nucleotides from the star point of primers CRYP*rbc*L2R (Hoef-Emden *et al.*, 2005).

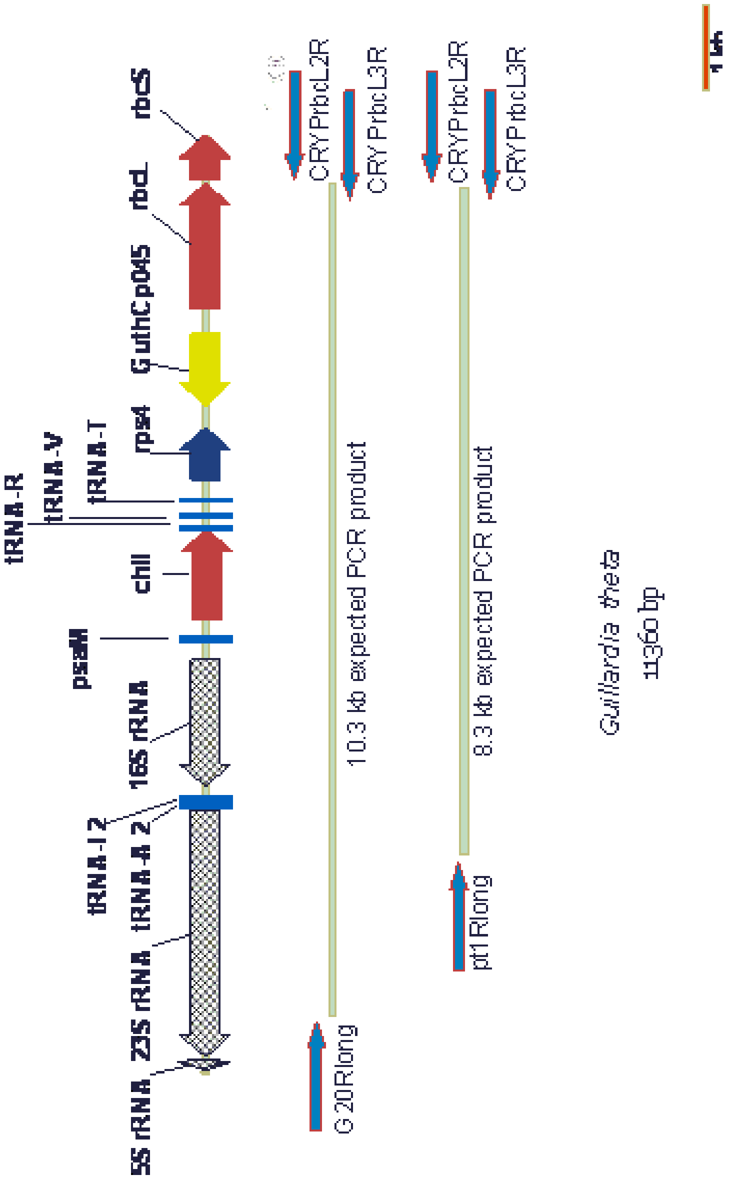


Fig. 2.6.1.1: *Illustration the approximately positions, manner combinations of long PCR primers and the expected long PCR products using* G. theta *plastome as reference.*

2.6.2 *Optimize the long-range PCR protocol and screen the primers*

Even though long-PCR approach has been using widely, optimization still needed to be done, at least in this study. MasterAmp^{TM} Extra-long PCR kit (Epicentre) was already prepared with 9 different Master Premix (numbered PreMix 1 to PreMix 9) containing a buffered salt solution with nucleotides, Mg^{2+} and enhancer (with betaine). To handle the kit, two rounds of reactions had to be run: the first round for searching the appropriate buffer(s) and the second one for optimizing PCR conditions. The component for one reaction was set up as the manufacturer's recommend but slightly modified.

For this purpose, total genomic DNA of strain CCAC 0031 was used as template; four couples of primers were employed from the combinations of two long forward primers and two long reverse primers:

- G20Rlong – CRYP*rbc*L2R, expected size: 10.3 kb
- G20Rlong – CRYP*rbc*L3R, expected size: 10.3 kb
- pt1Rlong – CRYP*rbc*L2R, expected size: 8.3 kb
- pt1Rlong – CRYP*rbc*L3R; expected size: 8.3 kb

Two different PCR programs were set up, one of which has 20 cycles and the other has 30 cycles of amplification (LONGPCR-20 and LONGPCR-30 programs, respectively). For each program, four sets of PCR were done for four couples of primers. In each set, nine different buffers were tested. The size of PCR products was expected around 8 – 10 kb depended on primer pair combination.

The testing results showed that buffer 1, 4 and 7; primers G20Rlong, pt1Rlong and CRYP*rbc*L2R performed better than others in both PCR programs.

Buffer 4, primer pair pt1Rlong – CRYP*rbc*L2R, and PCR program with 20 cycles (LONGPCR-20 program) were chosen as the optimal procedure and be used for further works.

Component for each PCR reaction:

Forward primer (100 pmol/µl)	1.5 µl
Reverse primer (100 pmol/µl)	1.5 µl
DNA template (100-150 ng/µl)	2-2.5 µl
MasterAmp Extrac-long DNA polymerase Mix:	0.5 µl
MasterAmp Extrac-long DNA 2X PreMix (buffer)	12.5 µl
Water	up to 25 µl

LONGPCR-20

Initially denature the template: 94^{o}C for 1’
Followed 20 cycles:
Denature at 98 ^{o}C for 20”
Anneal at 56 ^{o}C for 1’
Extend at 69 ^{o}C for 9’
Ended with 69 ^{o}C for 5’

LONGPCR-30

Initially denature the template: 94^{o}C for 1’
Followed 30 cycles:
Denature at 98 ^{o}C for 20”
Anneal at 56 ^{o}C for 1’
Extend at 69 ^{o}C for 9’
Ended with 69 ^{o}C for 5’

Reactions were run in an MWG biotech thermal cycler and kept at 4 ^{o}C overnight after amplification. The long PCR products were purified by DNeasy plant Mini Kit (250) according to manufacturer’s guide and stored in deep cold until use.

2.6.3 Verify the PCR products

Four cryptophyte strains CCAP 977/2a, CCAC 0113, CCAC 0109, CCAC 0031 were sampled randomly to amplify the 10 kb fragment between 23S rRNA and *rbc*L gene as done for CCAC 0031 in previous step. The results showed that the length of fragments vary from 8 kb to 10 kb as expected.

To verify the PCR products, two sequencing primers 1040R-700 (personal information from Dr. Birger Marin) and 1046R-700 (Hoef-Emden *et al.* (2005)) that were specific for 16S rRNA and *rbc*L gene, respectively, were employed to read these.

The sequencing results confirmed that the 10 kb fragments were actually comprised 16S rRNA and *rbc*L gene in at each ends.

2.6.4 Amplify the plastid 23S **rRNA–*rbc*L** *fragment by optimal long-range PCR protocol*

Twenty-two cryptophyte strains, including 4 heterotrophic strains and 18 autotrophic strains in which four strains were not genus *Cryptomonas* as out-group, were sampled to amplify the plastid 23S rRNA–*rbc*L fragments by optimal PCR protocols obtained above. Because the fate of *rbc*L gene in colorless strain M1634 was unclear, strain M1634 was delayed until *rps*4 gene in cryptophyte group to be read; then the 16S rRNA – *rps*4 fragment was amplified instead of the 23S rRNA–*rbc*L fragment as its sisters

Even thought the long PCR protocol was optimized, it was still needed some minor modification when applied each examined strains, mainly increasing or decreasing the DNA concentration and choosing the appropriate buffer(s). As expected, these PCRs ran successfully in all examined cryptophyte algae.

The PCR products were purified by elution from gel agarose electrophoresis and stored in -20 ^{o}C until use.

2.6.5 Produce the smaller segment by BioTherm™ Taq DNA Polymerase

Ten kb fragments of 21 cryptophytes harvested from long-range PCR above could be applied directly in reading the sequences. However, using only MasterAmp^{TM} Extra-long PCR kit to produce a huge amount of materials was considered not efficient economically. It was necessary to amplify smaller fragments by BioTherm™ Taq DNA Polymerase.

Two overlapping fragments, one from 23S rRNA to tRNA-T (called fragment 1) and the other from tRNA-R – *rbc*L (called fragment 2), were amplified using intermediate primers. For designing the intermediate primers, the tRNA-R (UCU) and tRNA-T (UGU) genes, located near the central of 23S rRNA – *rbc*L fragment according to *G. theta* plastome map, were chosen and named R_Rev and T_For, respectively.

Fragment 1 (around 4.5 kb) was produced by a combination of pt1Rlong and T_For run by PCR program set up named THD_PCR6 while fragment 2 (around 4 kb) was from combination of R_Rev and CRYP*rbc*L2Rlong primers run by THD-PCR4 program (see THD_PCR4 parameters as above).

As transcribed direction of both tRNA-R and tRNA-T to be opposite with *rbc*L gene, R_Rev, though its name, they worked as *forward* primer when combined with CRYP*rbc*L2R and similarly, T_For played its role as *reverse* primers when paired with pt1Rlong or/and G20Rlong primers (see Fig. 2.6.5.1).

The short PCR products were selected by visualization in agarose with ethilium bromide, then purified by QiaGene kit and stored in -20 ^{o}C until use.

THD-PCR6

Initially denature the template: 96^{o}C for 3'
Followed 30 cycles:
    Denature at 95 ^{o}C for 1'
    Anneal at 52 ^{o}C for 2'
    Extend at 68 ^{o}C for 4'
Ended with 68 ^{o}C for 5'

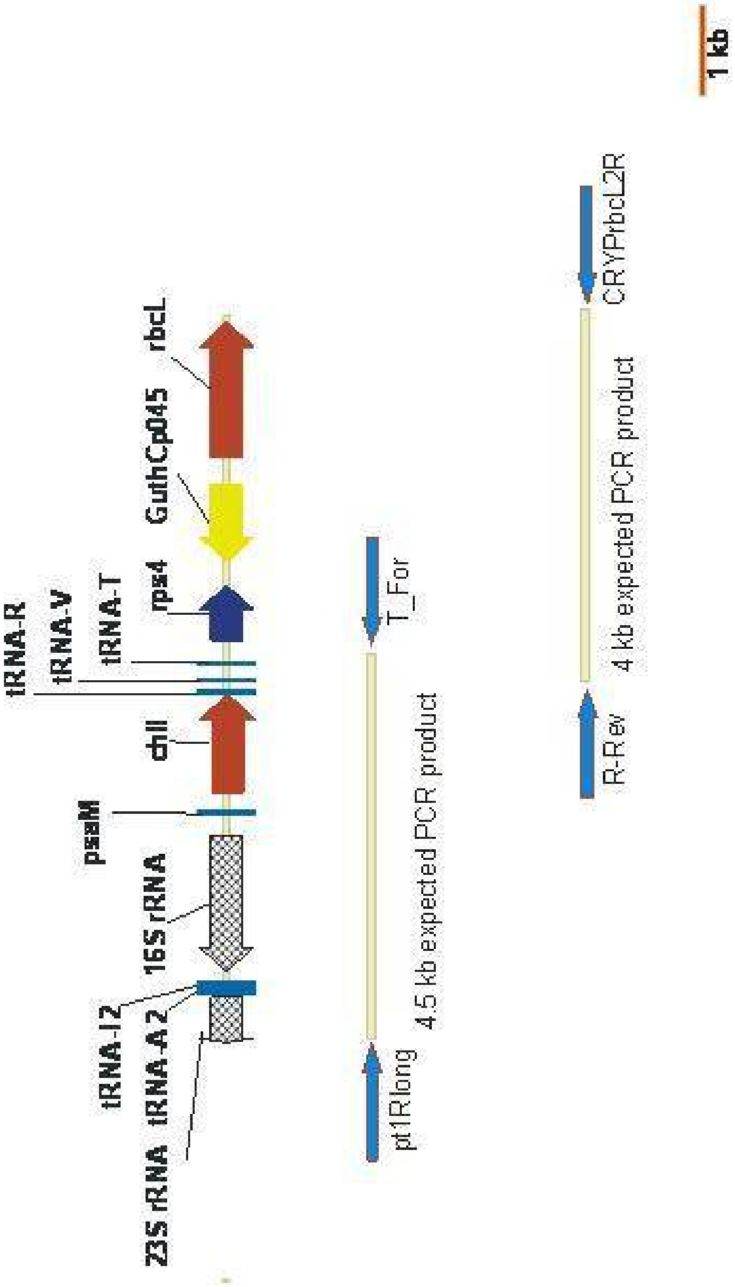


Fig. 2.6.5.1: *Illustration the approximately positions, manner combinations of PCR primers and the expected short PCR products using* G. theta *plastome as reference.*

2.6.6 Sequence the short PCR products (4kb) by primer-walking approach and using automated ABI3730 sequencer

To sequence double strands of the short PCR products obtained just above, BigDye Terminator v1.1 Cycle seqeucing kit 1000 reactions/unit (Applied Biosystems, Cat. N = 4337451) and ABI3730 automated sequencer were employed.

The lab-workings were begun at two ends of each fragment. In fragment 1, sequencing primer 16SH4Rev, a newly constructed primer that bound to the helix 4 domain of 16S rRNA, and primer T_For were used. Similarly, in case of fragment 2, R_Rev primer and primer *rbc*LStart_Rev, a newly designed sequencing primer that bound at approximate 120th position of *G. theta rbc*L, were employed.

The reactions were taken place on 96-well plates, with compositions for 1 reaction:

BigDye Buffer (5X): 0.5 µl
Primer (10 pmol/µL): 0.2 µl
Template: x µl (to enough 150-300 ng DNA)
Water: y ml to enough 2.5µl in total volume

<u>Seq-ABI1</u> program for sequencing

Initially denature the template: 94 °C for 2’
Followed 30 cycles:
    Denature at 94 °C for 2’
    Anneal at 40 °C for 30”
    Extend at 6 °C for 1’30’’
Ended with 60 °C for 6’

Reactions were run in an MWG biotech thermal cycler. After completion of PCR-sequencing, each of wells was applied for 17.5 µl of water before sent them to Sequencing Facility (Cologne Center for Genomics, University of Cologne) for reading by automated ABI 3730 system.

The draft sequences of the PCR products were imported to ChromasPro® version 1.41 (Technelysium Pty Ltd.) for proof-calling. The relative sequences were assembled into contigs using SeaView (Galtier *et al.*, 1996). The contigs from at least five examined cryptophytes were then collected for aligning to find the conserved domains near two ends of samples that would be employed to design next sequencing primers.

In most case, the sequencing primers were universal. However, in fact, some sequencing primers were inapplicable to, for example, *ycf*26, ORF403 genes in several Cryptophyta strains and *chlI* gene in all colorless strains. The *ycf*26 gene was detected in *R. salina* plastome but not in *G. theta* chloroplast genome; and it had not been planned to be sequenced in phylum Cryptophyta in the early state of the project since *R. salina* plastome had not been published. To overcome the problems, individual sequencing primers for specific cases were required. The universal and individual specific sequencing primers were listed and explained in detail in Table 2.6.6.1 and 2.6.6.2.

Table 2.6.6.1: The universal sequencing primers for cryptophyte plastomes 16S rRNA-*rbc*L fragments

	NAME	COMPOSITIONS	POSITION	DIRECTION
1	16SH14Rev	CGYAGGAGTCTGGGCCGTGTCTCAG	from the 290th to 314th position *G. theta* plastid 16S rRNA gene.	F(1*)
2	16SH4Rev	CGA CTT GCA TGT GTT AAG CAT ACC	from the 41th to 314th position *G. theta* plastid 16S rRNA gene	F
3	psaMF	YTRGCNATTMGWYTAGGAAC	from the 57th to 77th position of *G. theta* plastid *psa*M gene	F
4	psaMR	GCAACRAAWATTTGWGTRTCRCTAATC	from the 3rd to 29th position of *G. theta* plastid *psa*M gene	R(2*)
5	chl1F	AAGACCTCARYTWYTRGATMGATTTGG	from the 603th to 630th position of *G. theta* plastid *chl*I gene	F
6	chl1R1	CATHCCAAATCKATCYARWARYTGAGG	from the 607th to 633th position of *G. theta* plastid *chl*I gene	R
7	chl1R2	GCKGCRTRKGCTYTMGAAGCTC	from the 882th to 908th position of *G. theta* plastid *chl*I gene	R
8	R_For	GTC TAA KGG ATC AGG ACA GRR ACC TTC	from the 3th to 35th position of *G. theta* plastid tRNA-R gene	F
9	R_Rev	GGA TTC GAA CCT ACA TTA GAG ACT TAG	from the 35th to 61th position of *G. theta* plastid tRNA-R gene	F
10	V_For	AGCGGTAGAGCGCCTGCCTTAC	from the 13th to 34th position of *G. theta* plastid tRNA-V gene	R
11	T_For	GCT CAA TTG GTA GAG CAA CTG ATT TG	from the 8th to 33th position of *G. theta* plastid tRNA-T gene	R
12	T_for2	GCT CAA TTG GTA GAG CAA CTG ATT TGT	from the 8th to 34th position of *G. theta* plastid tRNA-T gene	R
13	T_Rev	GGA CTT GAA CCC GYA ACC KAC TGA TTA C	from the 33th to 60th position of *G. theta* plastid tRNA-T gene	F
14	rps4F	ATGTCYAGATAYAGAGGWGC	from the 1st to 20th position of *G. theta* plastid *rps*4 gene	F
15	rps4For2	GGA YAA TAT TGT DTT TAG RYT WGG NAT GGC	from the 276th to 305th position of *G. theta* plastid *rps*4 gene	F
16	rps4_For3	AAWGGNRTYATWGAAAGAGADTGGG	from the 529th to 553th position of *G. theta* plastid *rps*4 gene	F
17	rps4R	CGRCGMACKATYCTYAAYACKGCWCC	from the 16th to 41th position of *G. theta* plastid *rps*4 gene	R
18	rps4Rev2	GCN GGT ATW GTR GGN GCC ATN CC	from the 298th to 320th position of *G. theta* plastid *rps*4 gene	R
18	rps4Rev3	CCN GTR GAW CCT TTW ATT CTT CTW GC	from the 217th to 242th position of *G. theta* plastid *rps*4 gene	R
20	rps4_Rev4	CCTTTRCGWGARWARTRYTCHACWAC	from the 580h to 605th position of *G. theta* plastid *rps*4 gene	R
21	unknFor1	DRTRTADATNGGRKTTCCYTG	from the 430h to 450th position of CCAC 0086 plastid ORF403	F
22	unknRev1	TKKTTYCARGGAAMYCC	from the 440th to 456th position of CCAC 0086 plastid ORF403	R
23	rbcL_140R	TTTGAGGHGTCATACGGAACATWGC	from the 86th to 108h position of *G. theta* plastid *rbc*L gene	R
24	rbcLStart_Rev	GYT RTA RTC MGC ATC CCA RTA RCC C	from the 78th to 106th position of *G. theta* plastid *rbc*L gene	R

(1) F : Forward from 16S rRNA to rbcL*

(2) R: Reverse from rbcL to 16S rRNA*

Table 2.6.6.2: The specific sequencing primers for some Cryptophyta strains to read the plastomes 16S rRNA-*rbc*L fragments

	NAME	SEQUENCES	POSITION	DIRECTION
Specific primers for *ycf*26 and ORF403 gene of CCAC 0018 strain				
1	Rrev2_CCAC18	ATGATGAGTAGTGATGTTCGC	from the 52th to 72th positions of its *ycf*26	F (1*)
2	Tfor3_CCAC18	GATTACTTGCAAGATTATCTCCATC	from the 846th to 870th positions of its *ycf*26	R (2*)
3	UnFor2_CCAC0018	GTTATTGGTCCTTCATCGCGAGTC	from the 716th to 739th positions of its ORF403	R
4	UnRev_CCAC0018	CCTTAATTAATCAGCTGTCATCCGGCC	from the 1068th to 1094th positions of its ORF403	R
Specific primers for *ycf*26 gene of CCAC 0031 strain				
5	Rrev2_CCAC31	CCTTCTAACCTACTAAGTAGGTTCG	from the -155th to -131th positions of its *ycf*26	F
6	Tfor3_CCAC31	TGCCCTATCAGATCTGAGTTTCC	from the 649th to 671th positions of its *ycf*26	R
7	CCAC0031_Tfor4	CCTACTTAGTAGGTTAGAAGGATCC	from the -159th to -135th positions of its *ycf*26	R
Specific primers for ORF403 gene of CCAC 0113 strain				
8	UnknFor_CCAC0113	AGGCCTCGTTTAAGTCTAGTG	from the 510th to 530th positions of its ORF403	F
9	UnFor2_CCAC0113	CAGCGACATATGGATCCCATCGC	from the 921th to 943th positions of its ORF403	F
10	UnknRev_CCAC0113	CACTAGACTTAAACGAGGCC	from the 511th to 530th positions of its ORF403	R

Table 2.6.6.2 (continued): The specific sequencing primers for some Cryptophyta strains to read the plastomes 16S rRNA-*rbc*L fragments

Specific primers for *chl*I and ORF403 gene of M2452 strain				
11	chlIFor_M2452	TCCATTCGCAGCTATCGTAGGG	from the 39th to 60th positions of its *chl*I	F
12	16S-chlIRev_M2452	GTACACATAGCAATTCTATTTAGT	from the -18th to 6th positions of its *chl*I	F
13	UnFor1_M2452	GCTCATCATCAGAGTATGATCCTACCG	from the 390th to 416th positions of its ORF403	F
14	UnRev1_M2452	CAGAGTCCGTACGCTCTGATATAGG	from the 863th to 887th positions of its ORF403	R
Specific primers for *psa*M, *ycf*26, ORF403 and *rbc*L genes of CCMP 0704 strain				
15	psaMF1_CCMP704	GTGATACTCAAATTTTCCTAGCCC	from the 8th to 31th positions of its *psa*M	F
16	CCMP704_Rrev2	ACCTTATCCTTCGATCTGGAGC	from the 181th to 202th positions of its *ycf*26	F
17	CCMP704_Tfor2	TAGTAGCGGTTACCTGACTTATC	from the 594th to 616th positions of its *ycf*26	R
18	UnFor1_CCMP704	CCATTGCCTTAATTGCATCTTCG	from the 309th to 331th positions of its ORF403	F
19	UnRev2_CCMP704	CTAGCGACGAAGATGTGGTATTGG	from the 999th to 1032th positions of its ORF403	R
20	rbcL800F_704	CGGTATCGAAGGTGTTAACCGTGC	from the 672th to 695th positions of its *rbc*L	F

Table 2.6.6.2 (continued): The specific sequencing primers for some Cryptophyta strains to read the plastomes 16S rRNA-*rbc*L fragments

Specific primers for *ycf*26, ORF403 genes of CCMP 0443 strain				
21	CCMP_0443_Rrev2	CTAAGTGGACTCTACTATTTACG	from the -26th to -4th positions of its *yf*26	F
22	CCMP_0443_Tfor2	CATAATTACCCTTGATCTTAGG	from the 355th to 376th positions of its *ycf*26	R
23	UnFor1_CCMP0443	CTTGTACAGATATTTCCGAATCCG	from 362th to 385th positions of its ORF403	F
24	UnRev2_CCMP0443	CCCTGTAGTTCAATTACTAGAGG	from the 1107th to 1129th positions of its ORF403	R
Specific primers for *chl*I and *ycf*26 genes of SAG 980-1 strain				
25	SAG_980-1_Tfor2	CCTAGTTGTTTCTCGATTTGC	from the 163th to 183th positions of its *ycf*26	R
26	SAG_980-1_Rrev2	CGAGAAACAACTAGGTTGGTACGG	from the 169th to 192th positions of its *ycf*26	F
27	chlR3_980_1	GTCATCCTGGCCTACAATTGCGG	from the 50th to 72th positions of its *chl*I	R
Specific primers for *ycf*26 and ORF403 genes of M 1480 strain				
28	M1480_Rrev_2	TATCTGAATGAAGTTGGTG	from the 85th to 103th positions of its *ycf*26	F
29	M1480_Tfor2	GTGTAATCCCAATACCTGTGTC	from the 883th to 904th positions of its *ycf*26	R

Table 2.6.6.2 (continued): The specific sequencing primers for some Cryptophyta strains to read the plastomes 16S rRNA-*rbc*L fragments

30	UnFor1_M1480	CCCAGACTTTATATGCGTCATC	from the 321th to 343th positions of its ORF403	F
31	UnFor2_1480	CAT TAA GAC CTA CAT GTT CTG C	from the 645th to 666th positions of its ORF403	F
32	UnRev2_M1480	TACTTAGAACTATCTTATCAGG	from the -159th to -138th positions of its *rbc*L	R
Specific primers for *chl*I and *rps*4 genes of CCAC 0109 strain				
33	chlR3_CCAC0109	CCTCTATCTCCCATGATAATAACCCC	from the 118th to 143th positions of its *chlI*	R
34	CCAC0109_rps4For2	AGCTAATTTACCTACACACTTAGAG	from the 474th to 498th positions of its *rps*4	F
Specific primers for ORF403 gene of CCMP 0152 strain				
35	UnFor2_CCMP0152	CGACCGTGTAAATTGGTACCAC	from the 840th to 861th positions of its ORF403	F
36	UnRev2_CCMP0152	GATCTTCGTGAGGTCCAATGACTAG	from the -328th to -310th positions of its *rbc*L	R
Specific primers for ORF403 gene of M1092 strain				
37	UnFor2_M1092	CTCTTGGTGAAGCTGTGATAACTTC	from the 801th to 825th positions of its ORF403	F
38	UnRev2_M1092	CGATAATACATTCATTCAGAGCGG	from the 1013th to 1036th positions of its ORF403	R
Specific primer for ORF403 gene of CCAC 0064 strain				
39	UnRev2_CCAC0064	CCGCACAGATTCGTACTGAAAGG	from the -237th to -217th positions of its *rbc*L	R

Table 2.6.6.2 (continued): The specific sequencing primers for some Cryptophyta strains to read the plastomes 16S rRNA-*rbc*L fragments

Specific primers for ORF403 gene of CCAC 0056 strain				
40	UnFor2_CCAC0056	CGATGATGTCTCCAGACTTAGTGG	from the 930th to 926th positions of its ORF403	F
41	UnRev2_CCAC0056	CGGAAAAGACAGTCTTAGTCC	from the -246th to -226th positions of its *rbc*L	R
Specific primers for ORF403 gene of SAG 2013 strain				
42	UnFor1_2013	CGA TCT CGT GAA GGT CAG CTA CG	from the 534th to 556th positions of its ORF403	F
43	UnRev2_2013	GCT ATA TGT ACA TAC GTG CCC	from the -350th to -330th positions of its *rbc*L	R
Specific primers for ORF403 gene of SCAP K-416 strain				
44	UnFor1_K416	CGTAATATATAGATAGGTGTTCC	from the 389th to 411th positions of its ORF403	F
45	UnRev2_K416	CCTATGTGGAGGCTAGACCACAG	from the -380th to -358th positions of its *rbc*L	R

F$^{(1*}$) *: Forward from 16S rRNA to rbcL*

R (^{2*})*: Reverse from rbcL to 16S rRNA*

2.7 Annotate the genes

The nucleotide sequences of the 16S rRNA–*rbc*L fragment of each examined cryptophyte strains were transferred separately to Vector NTI (Invitrogene). Genetic code was set as Bacterial and Plant Plastid Code. ORF search was run in all six possible reading frames. Putative ORFs were translated to protein sequences for a PHI-BLAST search (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) with non-redundant protein sequences search set and then was annotated through comparison to those *G. theta* or/and *R. salina* CCMP 1319 plastome (Douglas & Penny, 1999; Khan *et al.*, 2007b). The tRNA genes were pointed out by using tRNAscan-SE search server at http://lowelab.ucsc.edu/tRNAscan-SE/.

2.8 Search for homologous protein by SMART server

The deduced protein sequence of each gene or ORF was submitted directly to SMART server (http://smart.embl-heidelberg.de/) to analyze the conserved domains and calculate E-value.

2.9 Align based on codon

The sequence collections were pre-aligned with SeaView (Galtier *et al.*, 1996), they then were imported to Selecton server for codon-based alignments (http://selecton.tau.ac.il/). Then, positions with codon-gaps in any of the sequences were omitted manually prior any phylogenies.

2.10 Analyze the phylogenetic issues

As first, the nucleotides or/and protein data sets were determined the most appropriate model of sequence evolution to use according to corrected Akaike Information Criterion (AICc) framework by jModeltest 0.1 and ProtTest1.4 with deactivated "+F" option, respectively (Posada, 2008; Posada & Crandall, 2001).

2.10.1 *For nucleotide data sets*

To construct the maximum likelihood trees, the date sets were uploaded to PhyML 3.0 (Guindon & Gascuel, 2003) using the proposed evolutionary model obtained from jModeltest 0.1.

Distances analyses were done by PAUP* 4.0b10 (Swofford, 2003), in which maximum likelihood parameter copied from jModeltest 0.1 and under minimum evolution. In case of data sets with heterogeneous base frequencies, the LogDet transformation was added into the phylogenetic process. The trees obtained from both these runs were inferred with neighbor-joining algorithm.

The PAUP* 4.0b10 was also employed to search for unweighted maximum parsimony trees using 10 random addition replicates in combination with heuristic tree search algorithm.

All of these runs, 1000 bootstrapped replicates were calculated.

For posterior probability analyses, MrBayes 3.1.2 (Ronquist & Huelsenbeck, 2003) was used with highly recommended settings such as Dirichlet distribution for relative substitution rates and base frequencies, bounded uniform distribution for proportion of invariable sites and gamma shape parameter for the distribution of among site rate variation, uniform distribution for topologies, and exponential distribution for branch lengths; 1.000.000 generations (due to the limited PC power) with one cold and three heated chains; trees saving every 100 generations and burn-in was determined manually according to the sum-plot displayed in the screen. In case of the combined data set, the analyses were performed with a partitioned model approach i.e. the appropriate parameters were calculated separately for each DNA partitions.

2.10.2 *For amino acid sequences*

After obtained the suitable evolutionary model, the protein data sets were subjected to the phylogenies like those of nucleotide data sets.

The PhyML 3.0 was used again for constructing the maximum likelihood trees in which CpRev model (Adachi *et al.*, 2000) was applied in all cases.

The neighbor-joining trees were done by MEGA 4 (Kumar *et al.*, 2008) with JTT (Jones-Taylor Thornton) matrix, gamma shape setting to results of ProtTest, and 1000 replicates.

Like the nucleotide sequences, un-weighted maximum parsimony trees for protein data sets were done by the PAUP* 4.0b10 using 10 random addition replicates in combination with heuristic tree search algorithm.

Bayesian analyzes for protein sequences were inferenced with MrBayes 3.1.2 with basic parameters like those of nucleotide data set exception for amino acid model was set to CpRev model ((aamodel=fixed(cpRev)).

All phylogenetic trees were displayed by TREEVIEW (Page, 1996).

3. RESULTS

3.1 Search and sequence the Cryptophyta plastome *rbc*L gene

Eight-teen newly sequences of the *rbc*L gene of *Cryptomonas* strains were obtained in the first part of the project were presented in the Table 3.1.1. Of these, four sequences were from heterotrophic (colorless) strains and the remaining was from photoautotrophic (pigmented) strains.

Five strains of three different colorless lineages were examined in this study: the colorless strain M1634 represented one lineage; strain CCAC 0056, CCAP 977/1, M2180 and M2452 represented the other; and strain SAG 977-2f represented the last.

Unfortunately, the strain SAG 977-2f did not survive during this study and some PCR attempts using the available frozen total DNA to amplify its *rbc*L were unsuccessful. Therefore, strain SAG 977-2f was rejected from the study.

PCR amplifications of the *rbc*L gene were successful in most colorless strains, but strain M1634 repeatedly gave negative results. A possible explanation for the unsuccessful PCR in strain M1634 was PCR primers did not fit. The reason may be that *rbc*L was highly diverged, was a pseudo-gene or was completely lost. Therefore, the existence of *rbc*L in the colorless M1634 was unclear. The actual situation of the *rbc*L gene of colorless strain M1634 should be studied in detail by another suitable strategy such as sequencing its whole plastid genome.

Table 3.1.1: List of *Cryptomonas* strains were read *rbc*L sequence

	Strain	**Accession Numbers**
01	*C. oobovoidea* CCAC 0031 (M1094)	AM051221
02	*C oobovoidea* CCAP 979/46	AM051223
03	*C. curvata* CCAC 0006 (M0420)	AM051204
04	*C. borealis* CCAC 0113 (M1083)	AM051202
05	*C. borealis* SCCAP K-0063	AM051203
06	*C. gyropyrenoidosa* sp. *nov.* CCAC 0108 (M1079)	AM051206
07	*C. lundii* CCAC 0107 (M0850)	AM051207
08	*C. marssonii* CCAC 0086	AM051208
09	*C. marssonii* CCAC 0103 (M1475)	AM051209
10	*C. paramaecium* CCAP 977/1	AM051213
11	*C. paramaecium* CCAC 0056 (M1303)	AM051212
12	*C. paramaecium* M2452	AM051215
13	*C. paramaecium* 2180	AM051214
14	*C. pyrenoidifera* CCAP 977/61	AM051216
15	*C. pyrenoidifera* CCMP 0152	AM051217
16	*C. pyrenoidifera* M1077	AM051218
17	*C. tetrapyrenoidosa* M1092	AM051219
18	*C. tetrapyrenoidosa* NIES 279	AM051220

3.2 Phylogenetic analyses based one the cryptophyte plastid *rbc*L gene

Eighteen newly sequences of the *rbc*L genes of *Cryptomonas* strains read in the first stage of the project were then combined with other five newly sequenced *rbc*L genes (done by Dr. Kerstin Hoef-Emden) that also belong to *Cryptomomas* for phylogenies. The results of phylogenetic analyses were compared with those of nuclear rDNA (concatenated SSU rDNA, ITS2 and partial LSU rDNA) and nucleomorph SSU rDNA to infer the evolutionary history of genus *Cryptomonas*. The *rbc*L sequences also were subjected to examine the codon usage of two-fold degenerate NNY codons to deduce the differences of functional constraints and expression levels in this gene in the genus *Cryptomonas* (these works were done by Dr. Kerstin Hoef-Emden).

The results showed that the *rbc*L gene in the colorless C. *paramecium* and their closely relative photosynthetic *Cryptomonas* had increased their evolutionary rates significantly. The shift from NNC to NNU in two-fold degenerate NNY codon in *rbc*L gene was recorded in the heterotrophic *Cryptomonas* and their closely relative photosynthetic *Cryptomonas*. The loss of photosynthetic activities in the colorless *Cryptomonas paramaecium* strains were explained by some possibilities. Detail results and discussion were published BMC *Evolutionary Biology* 2005;**5**:56.

In the second stage of the project, five newly sequenced *rbc*L gene (M2488, CCA0018, SAG 2013, M1480 and SCCAP K416) were harvested. They then were added into the previously *rbc*L data set for re-run phylogenies. The newly rooted maximum likelihood tree of cryptophyte *rbc*L sequences (51 taxa of 972 positions) was congruent with the tree published previously (46 taxa of 990 positions). The new adding *C. phaseolus* SAG 2013 grouped with *C. curvata* CCAC 0006 and CCAC 0080 but not always supported. Meanwhile, *C. loricata* M2088 attached with *C. commutata* CCAC0109 with moderate support (53/56/70/55/0.96; Fig. 3.2.1).

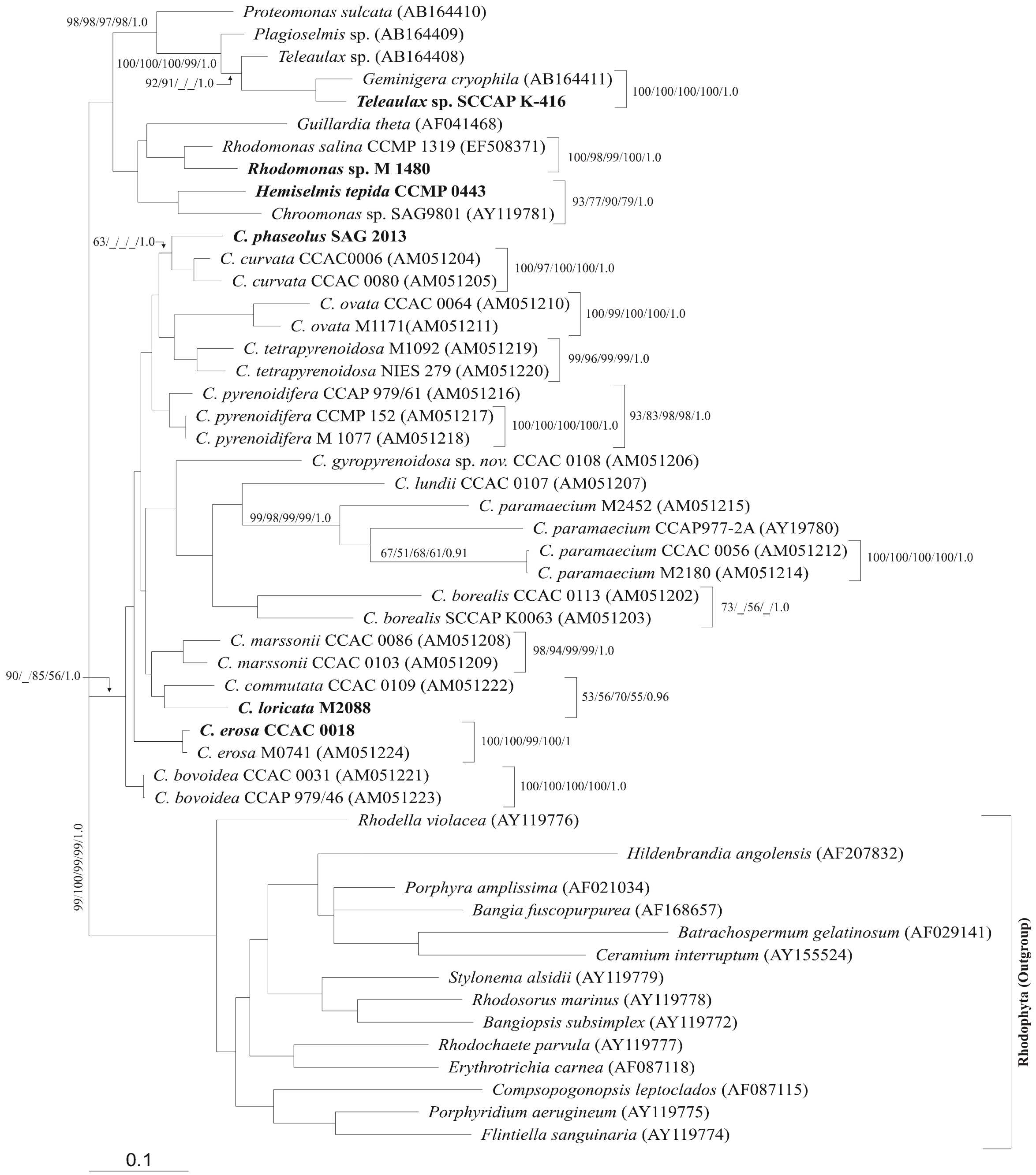

98/98/97/98/1.0
Proteomonas sulcata (AB164410)
Plagioselmis sp. (AB164409)
100/100/100/99/1.0
Teleaulax sp. (AB164408)
92/91/_/_/1.0
Geminigera cryophila (AB164411)
Teleaulax sp. SCCAP K-416
100/100/100/100/1.0
Guillardia theta (AF041468)
Rhodomonas salina CCMP 1319 (EF508371)
Rhodomonas sp. M 1480
100/98/99/100/1.0
Hemiselmis tepida CCMP 0443
Chroomonas sp. SAG9801 (AY119781)
93/77/90/79/1.0
63/_/_/_/1.0
C. phaseolus SAG 2013
C. curvata CCAC0006 (AM051204)
C. curvata CCAC 0080 (AM051205)
100/97/100/100/1.0
C. ovata CCAC 0064 (AM051210)
C. ovata M1171(AM051211)
100/99/100/100/1.0
C. tetrapyrenoidosa M1092 (AM051219)
C. tetrapyrenoidosa NIES 279 (AM051220)
99/96/99/99/1.0
C. pyrenoidifera CCAP 979/61 (AM051216)
C. pyrenoidifera CCMP 152 (AM051217)
C. pyrenoidifera M 1077 (AM051218)
100/100/100/100/1.0
93/83/98/98/1.0
C. gyropyrenoidosa sp. nov. CCAC 0108 (AM051206)
C. lundii CCAC 0107 (AM051207)
C. paramaecium M2452 (AM051215)
99/98/99/99/1.0
C. paramaecium CCAP977-2A (AY19780)
67/51/68/61/0.91
C. paramaecium CCAC 0056 (AM051212)
C. paramaecium M2180 (AM051214)
100/100/100/100/1.0
C. borealis CCAC 0113 (AM051202)
C. borealis SCCAP K0063 (AM051203)
73/_/56/_/1.0
C. marssonii CCAC 0086 (AM051208)
C. marssonii CCAC 0103 (AM051209)
98/94/99/99/1.0
90/_/85/56/1.0
C. commutata CCAC 0109 (AM051222)
C. loricata M2088
53/56/70/55/0.96
C. erosa CCAC 0018
C. erosa M0741 (AM051224)
100/100/99/100/1
C. bovoidea CCAC 0031 (AM051221)
C. bovoidea CCAP 979/46 (AM051223)
100/100/100/100/1.0
Rhodella violacea (AY119776)
99/100/99/99/1.0
Hildenbrandia angolensis (AF207832)
Porphyra amplissima (AF021034)
Bangia fuscopurpurea (AF168657)
Batrachospermum gelatinosum (AF029141)
Ceramium interruptum (AY155524)
Stylonema alsidii (AY119779)
Rhodosorus marinus (AY119778)
Bangiopsis subsimplex (AY119772)
Rhodochaete parvula (AY119777)
Erythrotrichia carnea (AF087118)
Compsopogonopsis leptoclados (AF087115)
Porphyridium aerugineum (AY119775)
Flintiella sanguinaria (AY119774)
Rhodophyta (Outgroup)
0.1

Fig. 3.2.1: *Rooted maximum-likelihood tree inferred from* rbc*L gene sequences (51 taxa of 972 positions). The evolutionary model (TIM2+I+G), - ln L= 16097.8824) were proposed by jModelTest 0.1 base.d on the Akaike information criterion. From left to right: maximum likelihood bootstrap/maximum parsimony bootstrap/neighbor-joining bootstrap/logdet transformation bootstrap/posterior probabilities (Bayesian analysis). Names in bold: newly sequenced* rbc*L obtained in the second stage of the project; scale bar: substitution per site.*

3.3 Annotate the 16S rRNA–*rbc*L fragments

In 22 Cryptophyte strains examined, 15 strains were read from 16S rRNA to *rbc*L without gap; and seven strains contain gaps due to the lack of time.

3.3.1 Gene contents

As mentioned in the previous part, the *rbc*L gene of colorless strain M1634 was not obtained, the attempts applying couple of primers pt1Rlong and CRYP*rbc*L2R, therefore, were unsuccessful repeatedly. Fortunately, using other couples of primers pt1Rlong with T_For and R_Rev with *rps*4Rev4 to amplify the 16S rRNA–tRNA-T and tRNA-R–*rps*4 fragments released two overlapping short fragments that subsequently to be sequenced; hence, M1634 was read from one-third end of *chl*I gene to end of *rps*4. Using *G. theta* and *R. salina* CCMP 1319 plastome as reliable references, the gene contents and gene orders of the 16S rRNA–*rbc*L fragments obtained above were annotated.

<u>16S rRNA gene</u> is a non-coding RNA gene translating the genetic information held in DNA to make the component of the small prokaryotic ribosomal subunit (30S). The conservation of rRNA genes (including 16S rRNA) are extremely high in all cells, therefore they have been used widely in investigating the phylogeny such as to identify an organism's taxonomic group, calculate related groups, and estimate rates of species divergence (Hillis & Dixon 1991; Harris *et al.*, 1994). Except for M1634, CCAC 0107 and CCAC 0108 strains had not been read the 16S rRNA gene, all examined cryptophytes were sequenced the 5'-terminus region of 16S rRNA with the numbers of nucleotides obtained from 116 to 557. The initial aligning showed they were strictly conserved as expected. Therefore, they are expected to be a convincible resource that someone can utilize in future project related to 16S rRNA gene of Cryptophyta algae.

***psa*M gene** is a small gene coding for a peripheral, membrane- integral subunits about 3.5 kDa involving the formation of stable Photosystem I (PSI) trimers that found in Cyanobacteria (Naithani *et al.*, 2000). The roles of *psa*M in the PSI complex of eukaryotic algae and higher plants have not been well-established as well as no paper reports finding this protein in PSI, despite being present in the chloroplast genome (Fromme *et al.*, 2001; Nelson & Yocum, 2006). In cryptophytes, all *psa*M genes are 93 bp in size with only one exception for 90 bp in M1092. The absence of *psa*M from colorless strains is a new finding.

***chl*I gene** (synonym: *bchl*I for *b*acterio*chl*orophyl) is a gene encoding for a soluble protein (*chl*I protein) with molecular weights of 38 – 45 kDa. This protein combines with other two close relatively components (*chl*D and *chl*H) forming the multisubunit Mg-chelatase that catalyses the insertion of Mg^{2+} into protoporphyrin IX in the first unique step of the chlorophyll synthesis (Suzuki *et al.*, 1997; Valentin *et al.*, 1998; Bollivar, 2006). Recent studies also demonstrated that the accumulation of Mg-protoporphyrin IX product was both necessary and sufficient for regulation by retrograde signaling of a large number of nuclear genes encoding plastid products (Rodermel, 2001**;** Strand *et al*., 2003; Nott *et al.*, 2006). Like those of *G. theta* and *R. salina*, the start codon of examined Cryptophyta *chl*I gene is GTG, but changing to ATG triplet in case of CCAC0113 and three colorless strains. The *chl*I gene in most cryptophyte species are 1062 bp long but fall down to 1002, 1020 and 1041bp in all three colorless CCAP 977/2a, M2452 and CCAC 0056, respectively.

tRNA genes include tRNA-R (UCU), tRNA-V(UAC) and tRNA-T(UGU), the lengths of which are identical in all strains, 73, 72 and 73 bp respectively. Only a minor difference observed in strain CCMP 0704 of which tRNA-T is 74 bp in size instead of 73 kb as usual. All three tRNA genes form a cluster in *G. theta* plastomes but they are separated by *ycf*26 in between tRNA-V and tRNA-T in *R. salina* plastomes. At least two *C. erosa* CCAC 0018 and

C. oobovoidea CCAC 0031 containing the ycf26 gene inserted into tRNA-V and tRNA-T is reported in this study.

***rps*4 gene** is a gene encoding protein 4 of the small plastid ribosomal (Douglas & Durnford, 1990; Harris *et al.*, 1994). Exception for M1634 lacking five last nucleotides because of the sequencing error, all of *rps*4 genes obtained are uniform in length, 609 bp. The predicted proteins have 203 amino acid residues. The start and stop codons are ATG and TAA in most case, respectively; some strains such as M1092, CCMP 0152, CCAC 0108 and CCAC0006 terminated by TAG triplet.

***ycf*26** is a gene that encodes a hypothetical sensor-like histidine kinase (Douglas & Penny, 1999; Khan *et al.*, 2007b). The presences of ycf26 in *C. erosa* CCAC 0018, *C. obovoidea CCAC 0031,* and non-*Cryptomonas* strains are unexpected.

ORF403 is a gene encoding Tic-22-like protein, one of five protein families involving to the pre-protein translocation at the inner envelope membrane of chloroplasts (Kouranov *et al.*, 1998; Reumann *et al.*, 2005). Unlike *ycf*26, this ORF403 presents in all cryptophytes. However, both cryptophyte *ycf*26 and ORF403 are highly divergent in size and nucleotide compositions, making the universal sequencing primers designed to read them useless. That was the reason why more specie-specific primers had been newly constructed to full-fill the *ycf*26 and ORF403.

***rbc*L** is a gene that encodes ribulose-1,5-biphosphate carboxylase/oxygenase with a very important role in the photosynthetic Calvin cycle as the carbon dioxide fixating enzyme. This enzyme catalyzes at least two reactions, the carboxylation of D-ribulose 1,5-bisphosphate and the oxidative fragmentation of the pentose substrate. The former is the primary event in photosynthetic carbon dioxide fixation and the later is in the photorespiration process. In the chloroplasts, both reactions work simultaneously and competitively at the same active site. The *rbc*L is a much-conserved gene and it has been used extensively by phylogenetists to unravel the phylogenetic history of plants (Kellogg and Juliano, 1997; Spreitzer & Salvucci**,** 2002) as well as algae (Melkonian *et al.*, 1995). Exception for M1634 and SAG 977-2f, all examined cryptophytes were read the *rbc*L genes.

In overall view, all cryptophyte strains which 16S rRNA–*rbc*L fragment were sequenced in this study can be divided into three small groups following their gene contents:

- **Group 1**: including most *Cryptomonas* strains such as CCAC 0109, M2088, CCAC 0006, CCAC 00113, CCAC 0086 (Fig. 3.3.1), SAG 2013, CCMP 0152 and M1092, with the gene contents, gene orders and transcribed direction like those of *G. theta* plastome.

- **Group 2**: comprising all colorless strains CCAP 977/2A, M2452 (Fig. 3.3.2) and CCAC 0056 which are similar to those of group 1 but lack of *psa*M gene.

- **Group 3**: containing two *Cryptomonas* strains CCAC0018, CCAC 0031 (Fig. 3.3.3) and five non-*Cryptomonas* strains SAG 980-1, CCMP 0443, CCMP 0704, M1480 and SCCAP K-416 that have an additional *ycf*26 gene in between tRNA-V and tRNA-T gene as observed in *R. salina* CCMP1319 plastid genome.

Strain M1634 can be put into group two as its colorless sisters; strains CCAC 00107, CCAC 0108 have not been grouped yet due to lack of information. Even though its ORF403 was not read, strain SAG 980-1 could be clastified to group 1 in an assumption that ORF403 always exists in all cryptophyte algae. The illustrations of the gene contents and gene order for the 16S rRNA–*rbc*L fragment in these groups were in Appendix 1 to Appendix 20.

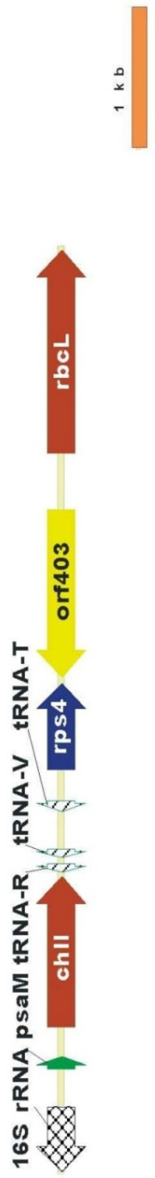


Fig. 3.3.1: *Illustration the gene content, gene order of the 16S* rRNA–*rbc*L *fragment in* C. marssonii *CCAC 0086 plastome.*

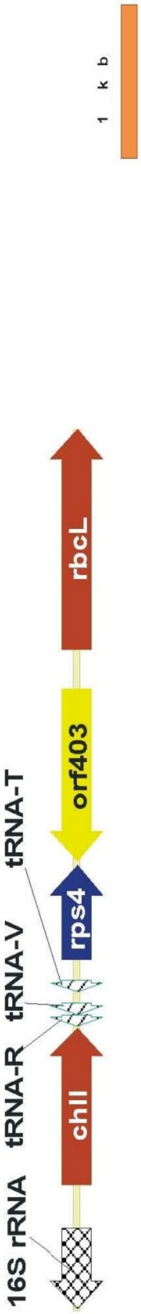


Fig. 3.3.2: *Illustration the gene contents, gene orders of the 16S* rRNA–*rbc*L *fragment in* C. paramaecium *M2452 plastome***.**

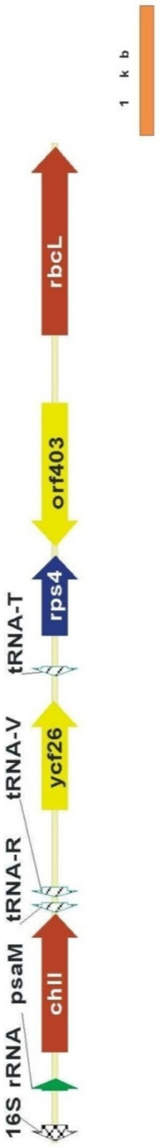


Fig. 3.3.3: *Illustration the gene content, gene order of the 16S* rRNA–*rbc*L *fragment in* C. *obovoidea CCAC 0031 plastome***.**

3.3.2 Overlapping genes

The overlapping gene was recorded in the last 43 bp of *rps*4 gene and ORF403 in colorless CCAC 0056. The overlapping phenomenon also observed in *R. salina* and *G. theta* plastomes such as between the pairs of *atp*D – *atp*F genes, *rpl*4 – *rpl*23 genes, *rpl*16 – *rpl*29 and ORF142 – ORF146 (Khan *et al.*, 2007b). The *rps*4 and ORF403 overlapped together is newly finding. The absence of *psa*M as well as the overlapping gene of *rps*4 and ORF403 found in colorless strains relate hypothetically to the plastid genome reduction in colorless *Cryptomonas*.

3.3.3 Compare the lengths of the fragments

Because the un-identical sizes of partial 16S rRNA genes, the distance from the starting points of 16S rRNA to those of *rbc*L genes were chosen and calculated to compare the 16S rRNA–*rbc*L fragment sizes among studied cryptophytes. Obviously, the presence or absence of *ycf*26 from the fragments was one of the reasons for the difference in lengths among examined cryptophytes. The lack of *ycf*26 and *psa*M genes in the 16S rRNA-*rbc*L fragments of the colorless group caused these fragments to be the shortest: 3667, 3748 and 4564 bp in sizes in CCAP 977/2a, M2452 and CCAC 0056, respectively. In the contrary, the dominantly longer 16S rRNA–*rbc*L fragments were in non-*Cryptomonas* and CCC 0018 and CCAC 0031 strains as they contain *ycf*26 genes. Surprisingly, two strains CCAC 0018 and CCAC 0031 hold the longest fragments: 6111 and 6074 bp in sizes, respectively. The differential value between CCAP 977/2a and CCAC 0018 were around 45%. This suggests that CCAP 977/2a and its colorless sisters possesses propably the smallest plastomes while CCAC 0018 and CCAC 0031 carried the largest plastomes among cryptophyte algae. Another suggestion is that the future whole plastome sequencing projects should focus in colorless CCAC 977/2a, M2452 and CCAC 0056 strains and photoautotrophic CCAC 0018 and CCAC 0031 strains together with another strain belongs to group 2 to compare the shortest, medium and longest plastomes among cryptophytes. See Table 3.2.3.1 for gene sizes in the fragments 16S rRNA–*rbc*L of studied cryptophytes.

Table 3.3.1: Gene sizes in the cryptophyte plastome 16S rRNA–*rbc*L fragments.

	Strain	**Sequenced length**	**Comparative length**	**SSU**	***psa*M**	***chl*I**	**tR**	**tV**	***ycf*26**	**tT**	***rps*4**	**ORF403**	***rbc*L**
1	CCAC 0109 (M0739)	6039 $^{(1*)}$	Over 5046	456 $^{(2*)}$	93	1062	73	72	NA	73	609	323$^{(3*)}$	1425 $^{(4*)}$ AM051222
2	CCAC 0018 (M0788)	7582	6111	252	93	1059	73	72	1332	73	609	1128	1220
3	M2088	6692	4962	506	93	1062	73	72	NA	73	609	1137	1226
4	CCAC 0031 (M1094)	7623	6074	116	93	1062	73	72	834	73	609	1101	1434 $^{(4*)}$ AM501221
5	CCAP 979/46	1441	Not calculated	-	-	-	-	-	-	-	-	-	1441 AM051223
6	CCAC 0006 (M0420)	4196	Not calculated	424	93	1062	73	72	NA	73	609	589$^{(3*)}$	-
7	CCAC 0113 (M1083)	7191	Over 5547	473	93	1062	73	72	NA	73	609	1461$^{(6*)}$	1310 $^{(4*)}$ AM051202
8	SCCAP K-0063	1322	Not calculated	-	-	-	-	-	-	-	-	-	AM015203
9	CCAC 0108 (M1079)	3690	Not calculated	-	-	-	-	-	-	-	609	1149	1436 $^{(4*)}$ AM051206
10	CCAC 0107 (M0850)	3310	Not calculated	-	-	-	-	-	-	-	609	550$^{(3*)}$	-
11	CCAC 0086 (M1473)	6510	4617	462	93	1062	73	72	NA	73	609	1173	1434 $^{(4*)}$ AM051208
12	CCAC 0103 (M1475)	1428	Not calculated	-	-	-	-	-	-	-	-	-	1428 AM051209
13	CCAC 0064 (M0847)	6288	Over 4819	216	93	1062	73	72	NA	73	609	1019$^{(3*)}$	1432 $^{(4*)}$ AM051210
14	CCAP 977/1	1430	Not calculated	-	-	-	-	-	-	-	-	-	AM051213

Table 3.2.3.1 *(continued):* Gene sizes in the cryptophyte plastome 16S rRNA–*rbc*L fragments.

15	CCAP 977/2A	5349	3667	250	NA	1002	73	72	NA	73	609	1152	1434 (5*) **AY119780**
16	CCAC 0056 (M1303)	6277	4564	261	NA	1041	73	72	NA	73	609	1212	1424 (4*) AM051212
17	M2452	5713	3754	529	NA	1026	73	72	NA	73	609	1116	1434 (4*) AM051215
18	M2180	1396	Not calculated	-	-	-	-	-	-	-	-	-	1396 AM051214
19	M1634	1660	Not calculated	-	-	464 (7*)	73	72	NA	73	605	-	-
20	SAG 2013	6333	4719	390	93	1062	73	72	NA	73	609	1143	1226
21	CCMP 0152	6408	4403	393	93	1062	73	72	NA	73	609	1172	1435 (4*) AM051217
22	CCAP 977/61	1420	Not calculated	-	-	-	-	-	-	-	-	-	1420 AM051216
23	M1077	1408	Not calculated	-	-	-	-	-	-	-	-	-	1408 AM051218
24	M1092	6406	4493	479	90	1062	73	72	NA	73	609	1170	1435 (4*) AM051219
25	NIES 279	1432	Not calculated	-	-	-	-	-	-	-	-	-	1432 AM051220
26	SAG 980-1	5404	Not calculated	249	93	1053	73	72	726	73	609	-	1095 (5*) **AY11971**
27	CCMP 0443	7373	5582	577	93	1053	73	72	738	74	609	1176	1217
28	CCMP 0704	6988	5503	260	93	1056	73	72	1053	73	609	1152	1226 (5*) **AB164410**
29	M1480	7245	5761	206	93	1059	73	72	969	73	609	1212	1226
30	SCCAP K-416	7536	5993	392	93	1065	73	72	705	73	609	1158	1252

***Sequenced length** is total length to be read in certain fragment; **comparative length** is the length of the distance from the starting point of 16S rRNA genes to those of* rbc*L genes. (-) the region was not read; (NA) the gene was not detected in the fragments; (^{1*}) The number of base pairs were read; (^{2*}) All 16S rRNA gene were read partial only at 3' terminus; (^{3*}) Only partial ORF403 was read; (^{4*})* rbc*L gene sequences was read and submitted to GenBank to get accession number in the first stage of this study (Hoef-Emden* et al.*, 2005); they were then full-filled the upstream gap afterward; (^{5}*)* rbc*L genes are available in the GenBank, they were re-read and full-filled the upstream gap in the second part of the study; (^{6*}) Strain CCAC 0113 has a unread region in noncoding space between ORF403 and* rbc*L gene; (^{7*}) one-third end of* chl*I gene was read.*

3.4 AT content and Identity characteristics

3.4.1 AT content

The AT contents were calculated for every genes in each strain (Table 3.4.1). Because the whole plastomes of examined strains are not yet available, the AT% of the 16S rRNA–*rbc*L fragments was used instead. The non-coding regions also were extracted for AT% content calculation. On the overall, the AT% of each genes in the colourless strains were not significantly different to the corresponding genes in the photosynthetic sisters. The *rbc*L, *rps*4, *chl*I and three tRNA genes held lower AT percentage than the whole fragments, while those of the *psa*M, *ycf*26 and ORF genes were higher; for example these values in strain CCAC 0018 are 57.5, 62.1, 62.8, 40.3, 66.7, 68.3 and 67.5% in comparison with its whole fragment AT% – 65.3%. These observations of AT% contents suggest that the *psa*M, *ycf*26 and ORF403 genes were being directed toward the genome compositional bias approach more strongly than *rbc*L, *rps*4, *chl*I and three tRNA genes.

3.4.2 Identity characteristics

The identity degree of nucleotide or/and protein sequences is one of the basic parameters to estimate the evolutionary distances between close relatives strains. Genes located in 16S rRNA–*rbc*L fragment were calculated the mean degree of sequence identity (Table 3.4.2).

For preparing the data set prior to analysis, collections of each *psa*M, *chl*I, *rps*4, *ycf*26, ORF403 and *rbc*L genes were done. The collection consisted of the newly read sequences plus those of *G. theta* and *R. salina* as references. Each collection was submitted separately to Selection server for codon-alignment. Then, positions with triplet-gaps in any of the sequences were omitted manually.

The data set of *psa*M gene contained 90 nucleotide positions of 18 taxa, none of which was colorless. In case of *chl*I gene, strain M1634 was rejected from the identity analysis because it composed only the last one-third of *chl*I sequence in length. The *chl*I dataset contained 21 taxa, three of them were colorless, and the sequence length was 963 nucleotide positions. Three tRNA genes were combined as one unit, consisting of 22 taxa of 217 characters. Strains CCAC 0117 and CCAC 0118 were absent from these three collections as they were not read from the *rps*4 to 16S-rRNA. Four colorless strains jointed with 20 photosynthetic cryptophytes to produce a collection of *rps*4 genes with 597 nucleotide characters. The ORF403 database had 20 strains and 345 nucleotide positions mainly distributed in the C-terminus, strains CCAC0107, CCAC0109 were out of the analysis due to their partial sequence readings. The largest collection was of *rbc*L, 30 strains of 972 nucleotide positions but only 3 colorless present while the shortest collection was of *ycf*26, only 8 strains and 462 nucleotides.

Then, the data sets were transferred to Vector NTI suite for calculating the identity values. The analyses were done in both nucleotide and protein characters with and without colorless strains.

The overall result showed that the identity values were very high in photosynthetic strains but they fell down when the non-photosynthetic strains were added to the comparison.

The lowest conservation was detected in *ycf*26 and ORF403 genes at both nucleotide and protein data sets in which their amplitude of identity values were very high: min value at 18% to 31% while max value at 77%. In ORF403 case, the identity values between photo and non-photosynthetic groups were not significantly different. Unsurprisingly, the three tRNA genes displayed the highest identity values in nucleotide data sets while *psa*M and *rbc*L proteins revealed the most conservation in protein comparison. Furthermore, the largest difference between colorless and non-colorless dataset was found in *chl*I gene.

Table 3.4.1: AT content of genes located between 16S rRNA and *rbc*L genes in examined Cryptophyta strains

	Strain	**16S rRNA -rbcL**	**16S rRNA**	***psa*M**	***chl*I**	**tR**	**tT**	***ycf*26**	**tV**	***rps*4**	**ORF403**	***rbc*L**	**noncoding-region**
01	CCAC 0109 (M0739)	63.1	49.6	66.7	60.6	52.1	40.3	-	53.4	62.6	65.5 *p	58.9	73.1
02	CCAC 0018 (M0788)	65.3	50.8	66.7	62.8	52.1	40.3	68.3	53.4	62.1	67.5	57.5	74.1
03	M2088	63.8	50.2	67.7	61.7	52.1	40.3	-	53.4	61.9	67.3	57.3	73.4
04	CCAC 0031 (M1094)	65.7	53.9	69.6	63.5	52.1	40.3	69.5	53.4	62.4	67.1	56.6	73.6
05	CCAC 0006 (M0420)	66.5	49.1	65.6	63.2	52.1	40.3	-	50.2	62.9	69.8 p	58.9	78.9
06	CCAC 0113 (M1083)	66.8	49.3	59.1	64.9	54.8	41.7	-	50.7	65.8	70.6	61.5	72.8
07	CCAC 0108 (M1079)	66.6	-	-	-	-	-	-	-	66.8	71.1	59	78.2
08	CCAC 0107 (M0850)	65.5	-	-	-	-	-	-	-	63.7	69.4p	60	74.5
09	CCAC 0086 (M1473)	65.7	47.9	68.8	63.1	52.1	40.3	-	53.4	64.7	71.8	58.5	78
10	CCAC 0064 (M0847)	66.7	52.3	66.7	64.5	52.1	40.3	-	53.4	64.2	66.8	59.6	77.7

Table 3.4.1 *(continued)*: AT content of genes located between 16S rRNA and rbcL genes in examined Cryptophyta strains

11	CCAP 977/2A	61	53.6	-	58.3	54.8	40.3	-	53.4	61.1	64.1	59.5	69.3
12	CCAC 0056 (M1303)	63.5	53.3	-	63.5	58.9	40.3	-	57.5	62.9	64.2	58.6	71.4
13	M2452	62.3	49.9	-	64.5	57.5	44.4	-	56.2	64.4	64.2	59.8	72.6
14	M1634	61.4	-	-	65.8***	50.7	47.2	-	54.8	59.9	-	-	71.3
15	SAG 2013	64.7	46.9	62.4	63.7	52.1	38.9	-	53.4	63.5	67.5	58.6	75.3
16	CCMP 0152	65.6	48.0	67.7	64.1	52.1	40.3	-	50.7	63.5	70.8	58.1	78.2
17	M1092	65.4	48.6	72.2	62.8	52.1	39.4	-	52.1	65	71.7	58	78.5
18	SAG 980-1	68.5	52.2	68.8	64	52.1	40.3	75.3	50.7	64.4	-	59.9	76.7
19	CCMP 0443	66.9	49.3	71	63.2	52.1	41.7	73	51.4	65.4	69.7	59.3	78.5
20	CCPM 0704	62.3	51.9	61.3	58.1	50.7	40.3	64.7	53.4	59.4	60.5	56.1	72.3
21	M1480	68.8	52.7	68.8	65.9	49.3	40.3	76.4	52.1	64.2	72.7	59	77.9
22	SCCAP K-416	66.3	47.2	68.8	63.4	49.7	40.3	69.4	52.1	63.7	70.7	58.7	75.4
23	*G. theta*	68.0***	50.2***	70	63.7	52.1	41.4	-	52.1	64.8	73.7	59.9	81.6
24	*R. salina*	64.0***	48.9 ***	68.8	63.6	50.7	40.3	71.1	52.1	63.7	70.4	58.3	79.2

(p): the genes were read partially, thus AT contents were not calculated

(-) The genes were not read or not presented.

Table 3.4.2: Comparison of identity values in non-photosynthetic and photosynthetic groups at both nucleotide and protein databases of genes in the Cryptophyta plastome 16S rRNA–*rbc*L fragments

	Number of Taxa	**Number of Nucleotide Positions**	**Nucleotide sequences**		**Protein sequences**	
			without colorless	with colorless	without colorless	with colorless
***psa*M**	18	90	68 – 88 %	-	70 – 100%	-
***chl*I**	21	960	71 – 88%	57 – 63%	84 – 95%	50 – 59%
3tRNA	22	217	94 -100%	89 – 95%		
***ycf*26**	8	462	18 – 57 %.	-	18 – 52%	-
***rps*4**	24	597	68 – 88%	64 – 73%	74 – 96 %	62 – 82%
ORF403	20	345	48 – 77%	45 – 50%	31 – 72%	31 – 61%
***rbc*L**	30	972	80 – 94%	77 – 85%	88 – 99%	86 – 90%

(-) not calculated due to the lack of these genes in colorless strains

3.5 Detect Shine-Dalgarno sequence, -35 and -10 boxes of genes in 16S rRNA–*rbc*L fragments

In this part of study, the inter-space sequences in between 16S – *psa*M, *psa*M – *chl*I, tRNA-T–*rps*4, *rps*4–ORF403 and ORF403–*rbc*L were isolated to search for the transcriptional and translational regulation elements, among which focusing to Shine-Dalgarno (SD) sequences, -35 box and -10 box as these elements play very important roles in translational and transcriptional regulation for gene expression (Hirose & Sugiura, 2004; Somanchi & Mayfield, 2004).

The results showed that the upstream region of *rbc*L gene was the easiest alignable region then followed by those of 16S rRNA–*psa*M and tRNT–*rps*4; the upstream of *chl*I gene was very high variant making it unalignable. Through the aligning process, the conserved domains in the upstream regions of most photosynthetic strains were easily recognizable while those of colourless, CCAC0107, CCAC0108 and CCAC113 strains were not.

Among the express regulating elements located in the upstream, the SD sequences were exposed obviously. They were found in the upstream region of *rbc*L, *psa*M and *rps*4 with AGGAG motif. Surprisingly, the *chl*I genes consisted of no SD sequence but with an adenine-rich cluster. The common core AGGAG was also used to search for SD sequence in ORF403 and *ycf*26 but no results returned.

Douglas *et al.* (1990) pointed out the SD sequence – AGGAGG – in *rbc*L of *G. theta*. In our *rbc*L data set, this formula was found to be highly conserved. However, some changings were observed in strain colourless and CCAC 0113 – the first A altered to C or G or T and the second A changed to G.

While the conservation of SD block in *rbc*L genes was very strict, the situation in *rps*4 genes was relax. The SD core sequences of *rps*4 were formulated as GGAGAA. The SD clusters in two colourless strain CCAP 977/2a and M1624, again, showed the difference in comparison with their sister – first uninterrupted A and ended with GTT.

The SD clusters of two strains CCMP 0152 and M1092 were recored with four or five G then ended with one A while those of strain CCAC 0006 and CCAC0086 ended with four A.

The lost of *psa*M gene as well as the compress of the genomes in colorless strains caused their inter-space between 16S rRNA and *psa*M/*chl*I became shorter. The SD sequences and other express regulation sites, therefore, were probably lost as well. In the remain photosynthetic strains, a conserved block in front of *psa*M genes - A(C)GGA(T)G(A)A(T/G) - can be assigned as SD sequences.

The identification of -35 and -10 boxes for *rbc*L dataset were based on the description from Douglas *et al.* (1990). The -10 box sequences displayed a high similarity in photosynthetic strains but not in colourless, CCAC0107, CCAC0108 and CCAC0113 strains. Douglas *et al.* (1990), pointed out the -35 box of *G. theta rbc*L gene (TTGAGT) was around 19 nucleotides distance to -10 box, however, the non coding regions in the upstream of *rbc*L in this study showed that this cluster could be inexact. Instead, at least two high-conserved clusters could be marked as potential -35 boxes. The conclusion of exactly position of these two regular elements is expected to be done by reverse genetics or any experiments in the future.

Due to the lack of information of the regulation boxes of *rps*4, *psa*M, *ycf*26, ORF403 and 16S rRNA genes for cryptophytes, the sequences of interested regular elements were identified by scanning the conserved blocks in the upstream regions of these genes by eyes. Two blocks, around 110 nucleotide from start triplet of *G. theta,* were assigned as prospect -35 and -10 boxes for *rps*4 genes. Three conserved blocks emerged as potential regulation sequences in the non-coding region of *psa*M while one unknown-function conserved domain was found in the upstream region of 16S rRNA. The finding for regulatory elements in ORF403 and *ycf*26 were unsuccessful.

Detail of conserved block, -35 box, -10 box and SD sequences in non-coding regions of 16S rRNA, *psa*M, *chl*I, *rps*4 and *rbc*L genes were illustrated in Fig. 3.5.1 to Fig 3.5.5, respectively.

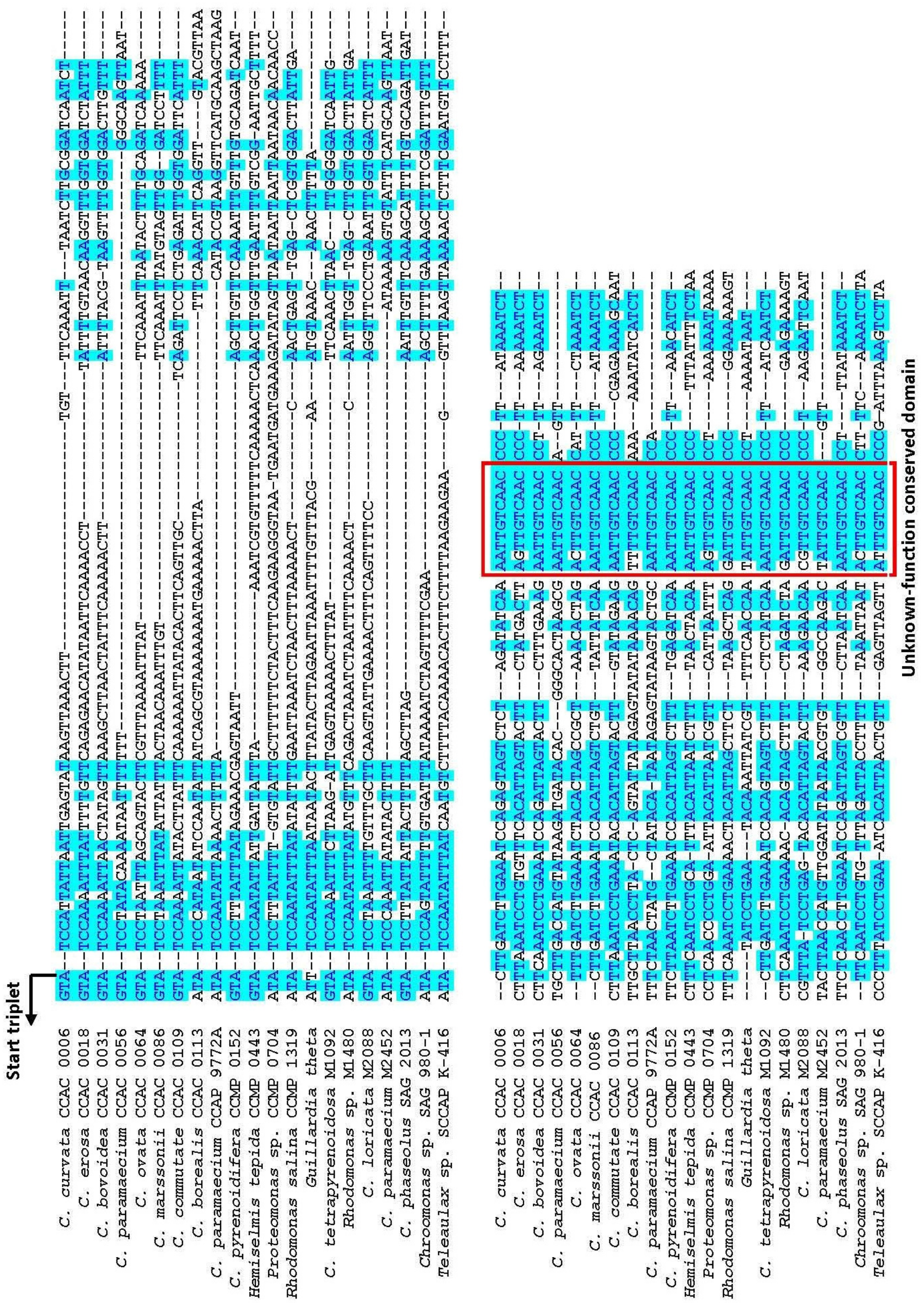


Fig. 3.5.1: *Alignment for 16S rRNA upstream region, one unknown-function conserved was found***.**

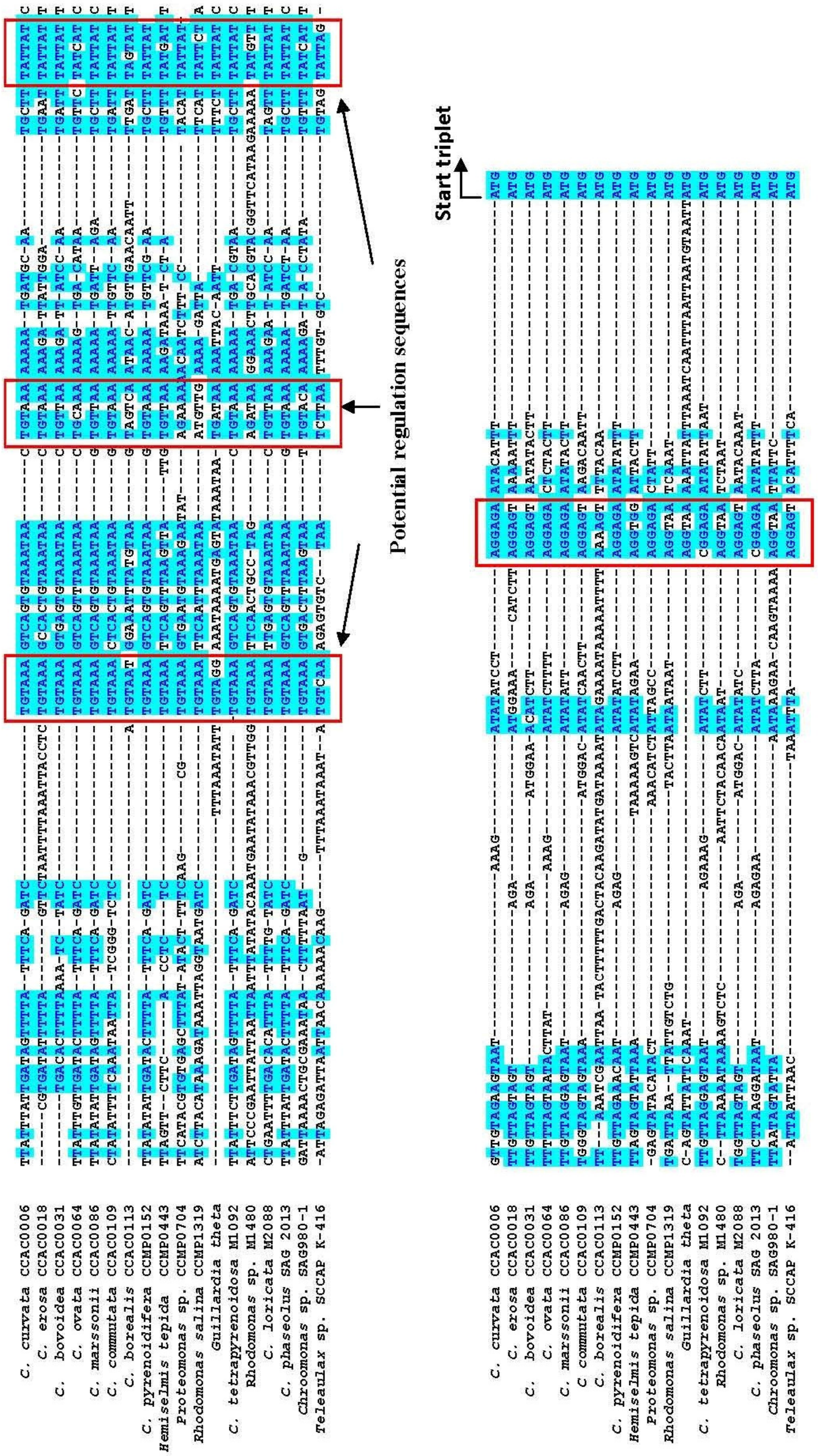


Fig. 3.5.2: *The non-coding region of* psa*M gene, one SD sequences and three potential regulation sequences were illustrated.*

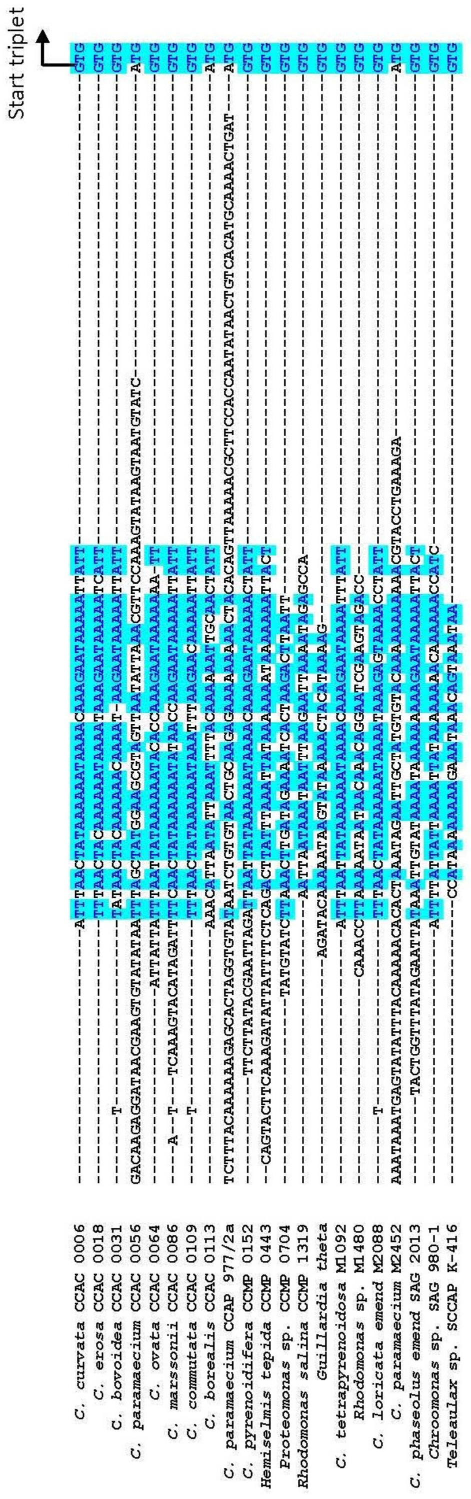


Fig. 3.5.3: *The upstream part of* chl*I gene, no SD sequence or regulation elements was detected, but an adenine-rich block emerged***.**

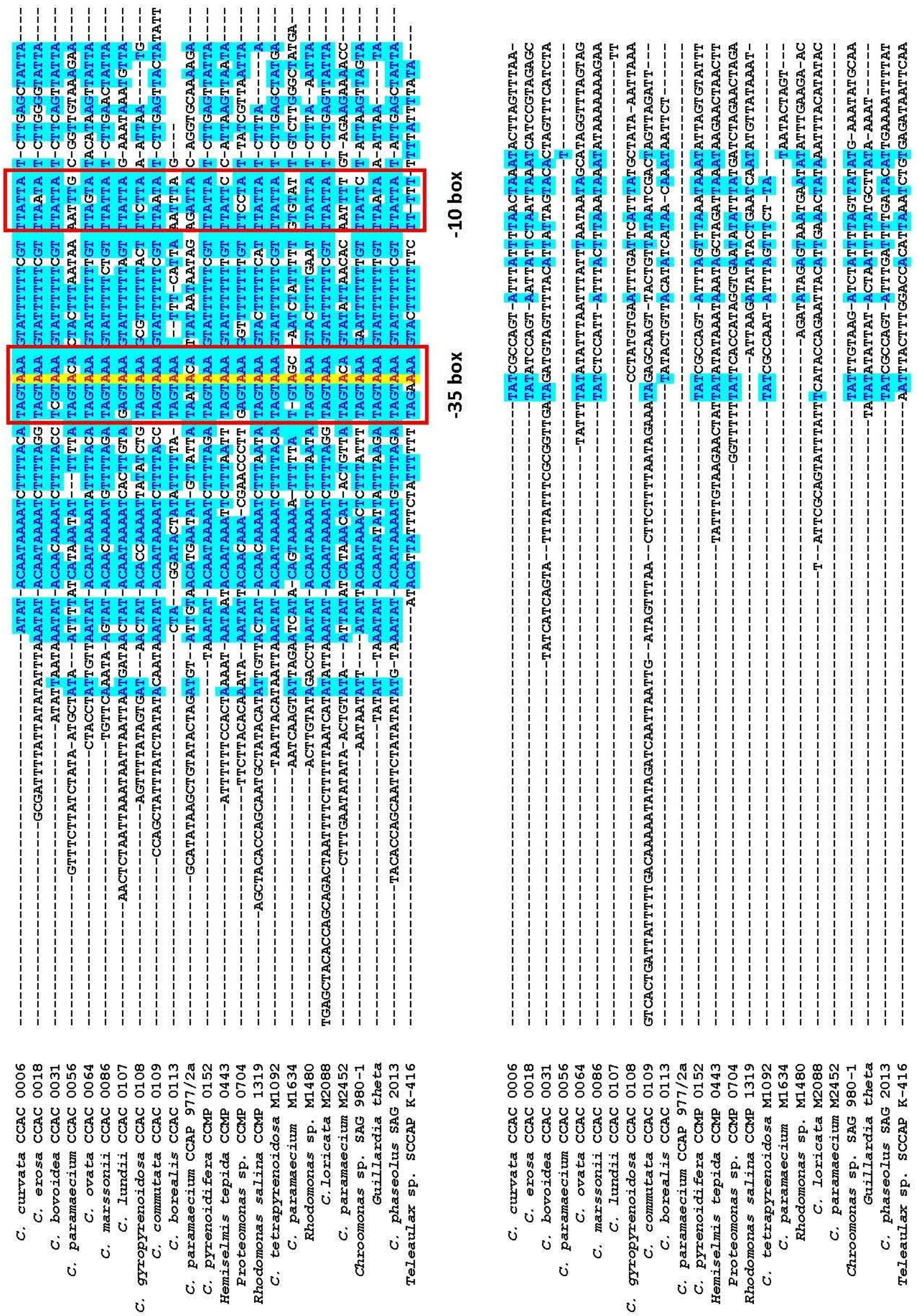


Fig. 3.5.4: *Alignment for upstream region of* rps*4 gene, the -35, -10 boxes and SD sequences were recognized*.

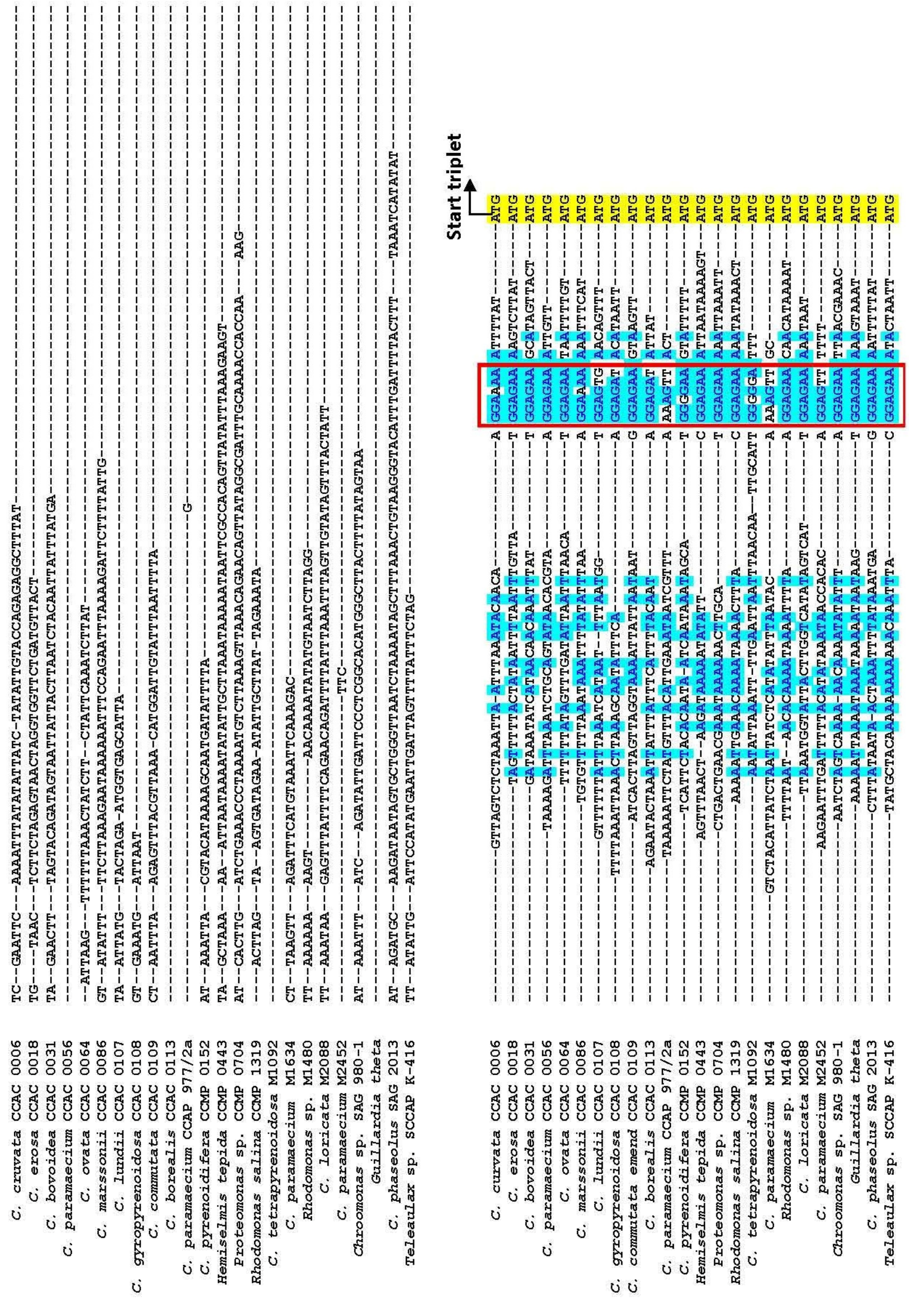


Fig. 3.5.4 *(continued): Alignment for upstream region of rps4 gene, the -35, -10 boxes and SD sequences were recognized.*

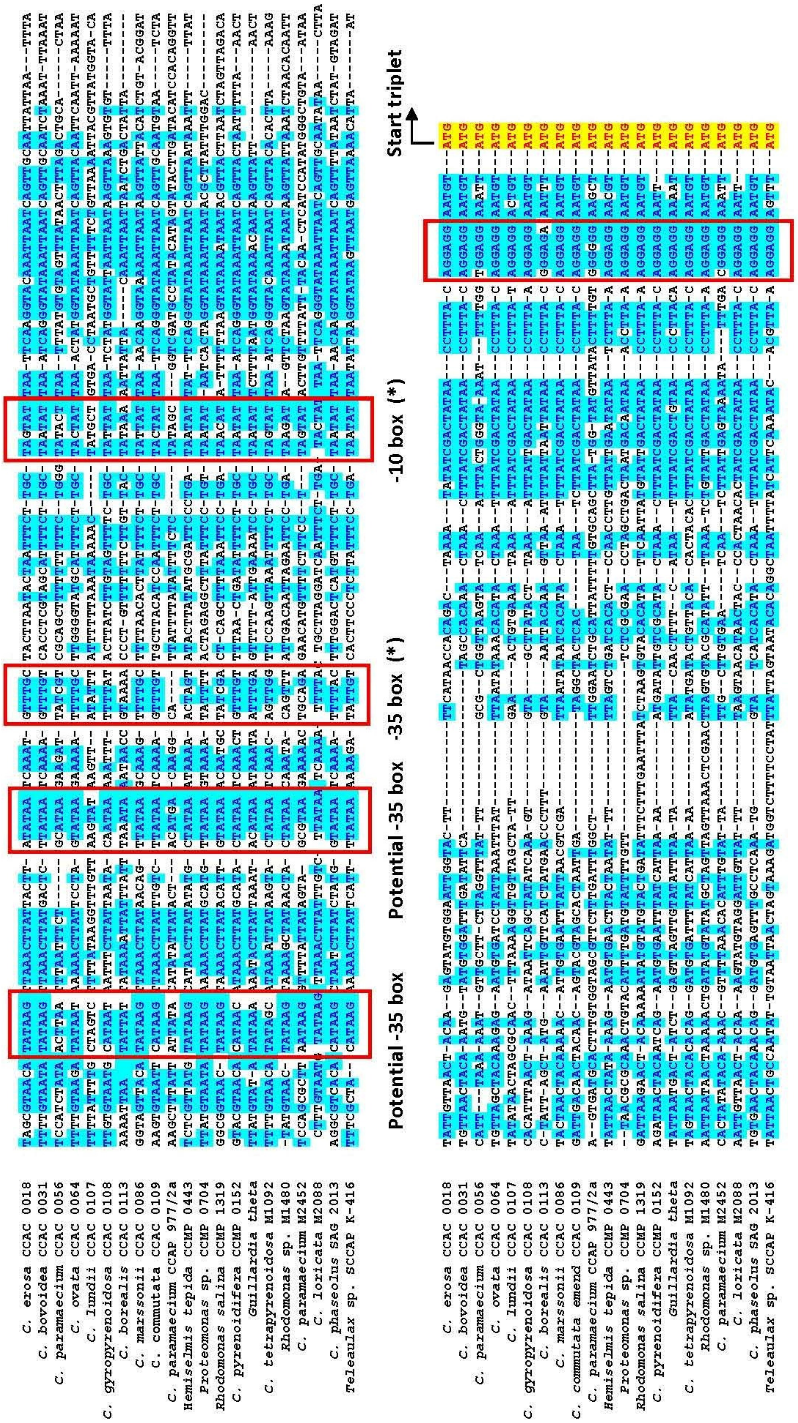


Fig. 3.5.5: *The non-coding region of* rbc*L gene, the SD sequences, -35 box and -10 box (marked with *) pointed out by Douglas* et al. *(1990) beside two new potential -35 boxes.*

3.6 *ycf*26 – Unexpected protein in *C. erosa* CCAC 0018 and *C. ovoidea* CCAC 0031

3.6.1 The distribution of ycf26 *in photosynthetic organisms*

The sequencing and annotating results showed that there are ORFs lying between tRNAt and tRNA-V in the 16S rRNA–*rbc*L fragments of several cryptophytes such as *C. erosa* CCAC 0018, *C. obovoidea* CCAC 0031, *Proteomonas* CCMP 0704, *Hemiselmis tepida* CCMP 0443, *Teleaulax* SCCAP K-416, *Rhodomonas* M 1480 and *Chroomonas* SAG 980-1. These ORFs were submitted to NCBI to search for the homologies; all the results obtained showed that these ORFs were similar with *ycf*26 of *R. salina* CCMP 1319. Surprisingly, the *ycf*26 gene was detected in *R. salina* plastome but not in *G. theta* chloroplast genome; and it had not been planned to be sequenced in phylum Cryptophyta in the early state of the project since *R. salina* plastome had not been published. The presences of ycf26 in *C. erosa* CCAC 0018, *C. obovoidea* CCAC 0031 and non-*Cryptomonas* strains are unexpected.

The *ycf*26 was originally named for a hypothetical chloroplast ORF with 1968 bp long that was found at first time in *P.purpurea* plastome (Reith & Munholland, 1993). It then was discovered in all genomes of Cyanobacteria and in some red-line plastomes such as rhodophyte, haptophyte, raphidophyte, but it was absent totally in green-line chloroplasts. The finding of *ycf*26 in only two strains *C. erosa* CCAC 0018 and *C. obovoidea* CCAC 0031 in this study also showed its non-universal distribution in species level. The *ycf*26 has been called under different names such as hik33 (sensor-like *hi*stidine *k*inase), chk33, *ycf*26 (hypothetical chloroplast protein), dfr (*d*ihydro*f*olate *r*eductase), dspA (*d*rug *s*ensor *p*rotein A), nblS (*n*on-*bl*eaching mutant *S*ensor), tsg1 (*t*ranscriptional *s*ensor *g*ene 1) depending on the pioneer researchers who gave the names (Ashby & Houmard, 2006); from now on, *ycf*26 will be used for short. Summary of distribution of *ycf*26 is in Table 3.6.1.

3.6.2 Describe the functional structure of ycf*26 protein*

Due to the major critical roles in biological activities of bacteria, *ycf*26 drew much attention from researchers. The *ycf*26 gene encodes a sensor-like histidine kinase (Hik), which functions were deeply investigated in bacteria, including Cyanobacteria. The sensor-like histidine kinase (Hik) is the common name for a protein super-class that perceive the extracellular environmental stimuli then transfer these signals to their partners (called *R*esponse *re*gulator – Rre) subsequently regulate a wide variety of cellular processes to response to these stresses. The sensor and regulator form a so-called two-component signal transduction system that is vitally important for cell survival and growth. Briefly, the cyanobaterial Hik33 has been proved as a "multi-stress" sensor, receiving many extracellular signals such as strong light, nutrient, chemical, temperature and osmotic stresses (Stock *et al.*, 2000; Suzuki, 2000; Mikami *et al.*, 2002; van Waasbergen *et al.,* 2002; Mikami *et al.,* 2003; Morrison *et al.*, 2005; Ashby & Houmard, 2006; Morici *et al.*, 2006; Kanesaki *et al.*; 2007).

Unfortunately, almost nothing has demonstrated the expression signals of *ycf*26 in both mRNA and protein levels in algae. Duplessis *et al.* (2007) planned to locate the position of *ycf*26 protein in chloroplast of *Heterosigma akashiwo* by monoclonal antibody, but no results have been published still now. The analysis for *ycf*26 in this part, therefore were based on the described cyanobacterial *ycf*26 (Morrison *et al.*, 2005). For convenience, the functional domains of *Synechocystic* PCC 6803 *ycf*26 is described here before steps forward in next parts.

The predicted *ycf*26 structure in *Synechocystic* PCC 6803 comprises 663 amino acids building seven functional domains. Two potential TMH regions (*t*rans-*m*embrane *h*elical) at 33 – 55 and 201 – 223 positions help the protein integrate into cell membrane. Inserted between two TMH regions is a 140 amino acid long segment (at 58 – 197 positions). This strictly conserved region in all cyanobacterial *ycf*26 is assumed to be putative *p*eriplasmic *s*ensor *d*omain (PSD), playing the role of a periplasmic sensor. Overlapping with the second TMH is the HAMP region (*H*istidine kinases, *A*denylyl cyclases, *M*ethyl binding proteins, *P*hosphatase) located at 200 –

269 positions which plays the role of transmitter, to transfer the extracellular signal from PSD domain to into Hik33 cytoplasmic regions. Beside the HAMP is the PAS domain (*P*er, *A*rnt, *S*imi or *p*eriod protein, *a*ryl hydrocarbon receptor nuclear translocator protein and *s*ingle-mined protein) positioned at 283 – 349. An assumption is that the combination of PSD and PAS domains enables the capability of *ycf*26 to receive a wide variety of extracellular stimuli. The two most important regions are HiKA (422 – 490 positions) and HATPase_c (535 – 656 positions) domains nearly the C-terminus of Hik33 polypeptides. The HiKA is responsible for dimerisation and ATP-dependent autophosphorylation of the histidine residue and subsequently transfer the phosphoryl group from the kinase to an aspartate residue of the response regulator by the catalysis of HATPase_c domain. Scattered distribution in both HiKA and HATPase_c domains are highly conversed boxes named H, N, F, G1, G2 and G3 following the main amino acids located in these boxes. The H box near the N-terminus contains the consensus amino acid array HELRT (432 – 436 positions) in which the histidine residue is phosphorylation site and the array to be considered as indicators for classifying of sensor kinases. The N (NLIGNS), G1 (ISDTGI), F (IEFREYR), G2 (GTGL) and G3 (only G) boxes are found closer to the C-terminus in a catalytic domain responsible for Mg^{2+} and ATP binding at position 546 – 551, 586 – 591, 601 – 607, 617 – 620 and 645 positions, respectively.

3.6.3 Search the functional domains of red-line plastid **ycf26**

Fifteen *ycf*26 protein sequences, including eight new cryptophyte *ycf*26, six other plastids and one from Cyanobacterium *Synechocystic* PCC6803, were submitted separately to SMART server for searching the functional domains.

The results obtained from SMART server showed that 14 plastid *ycf*26 could be grouped into two clades. Clade A included four *ycf*26 from Rhodophyta and one from Haptophyta *Emiliania huxleyi* in which their functional domain numbers were similar to those of Cyanobacteria. Clade B comprised eight strains of Cryptophyta and one of Heterokontophyta in which the average length was around 316 residues, nearly a half of those of primary plastids, containing mostly the HiKA and HATPase_c domains only.

SMART identified a very high accuracy (95% to 100%) in almost the functional domains of Cyanobacteria (PCC6803) *ycf*26 with except for PSD. It is possible that the PSD has not been archived in the SMART database. The domains correspond to the PSD in strains of clade A were reported as the intrinsic disorder regions; they therefore were annotated based on the aligned proteins obtained from Selecton server and according to description from Morrison *et al.* (2005). Observation the PSD of *ycf*26 proteins from plastids showed that they actually were not as conserved as *ycf*26 from Cyanobacteria. These suggested the regions that function like periplasmic sensors for the environmental signals seemed to be changed either to receive other specific stimuli for plastids or to degenerate their functions.

Five other putative functional domains were assigned for *ycf*26 from primary plastids and *E. huxleyi* with the support e-value were similar to those of PCC 6803.

In the opposite, the two trans-membrane helical, HAMP and PAS regions were not found in the *ycf*26 protein sequences of the truncated-*ycf*26 clade; only HiKa and HATPase_c domains existed but very weak E-value. SMART also pointed out so many intrinsic disorder domains distributed across the *ycf*26 protein sequences of this clade.

The presence of the array HELRT in the conserved H box helped to verify that the plastid *ycf*26 proteins actually belong to type one sensor. Among the conserved boxes, the H and N boxes were highly conserved while the F, G1, G2 and G3 were relaxed. The G1, G2, and F boxes were collapsed completely in strain CCAC 0031. The F box was nearly not recognized in strain CCAC 0018 and M1480 because of the strongly amino acid composition altering in these boxes.

Details of starting and ending points of each functional domain and E-values obtained from SMART server for 14-plastid *ycf*26 and one Cyanobacterium were listed in the Table 3.6.2.

Observing the *ycf*26 genes from CCAC0031 and CCAC0018 revealed an unusual difference in length (831 versus 1329) and start codon (CTG instead of ATG). It was doubtful that there was a relic of *ycf*26 exepctely in the upstream part of current CCAC0031 *ycf*26. ORF detection for the upstream region of CCAC 0031 *ycf*26 was done again. Fortunately, a small ORF (257 nucleotides) in the beginning of current CCAC 0031 *ycf*26 was found. This additional ORF was transferred to NCBI for BALSTX search. Surprisingly, it matched with a protein of *Tetrahymena thermophila* SB210 (gene ID 4512514) containing protein kinase domain. Meanwhile, the translated ORFs (+1, +2 and +3) were put into SMART server to detect functional domains. Again, it was interesting that two of these translated ORFs contained one or/and two trans-membrane regions without E-values while the other one held nothing.

The functional domains of *ycf*26 protein of cryptophytes and some random chosen Cyanobacterium were illustrated in Fig.3.6.1.

3.6.4 Reconstruct the phylogenetic trees

For the first view of the evolutionary history, the phylogenetic trees were conducted. To increase the number of taxon sampling, 14 red-line algae and 45 Cyanobacteria *ycf*26 sequences were jointed for a large dataset comprises 59 taxa (see Appendix 22 for detail of names and accession numbers of Cyanobacteria). The *ycf*26 collection was then submitted to Selecton server for codon-alignment. Alignable condon-based regions in nucleotides or/and protein sequences were extracted prior phylogenies.

The jModelTest 0.1 and ProtTest 1.4 proposed TVM + G + I and JTT + G (Posada, 2003) as the best fitting evolutionary for nucleotide and protein data set, respectively.

The red-line *ycf*26 always formed a separated clade in phylogenetic trees (Fig. 3.6.2 and 3.6.3). In both trees, their evolutionary rates were extremely high, of which the most significantly increased was *Teleaulax* sp. SCCAP K-416. There was a competition for the basal position in red-line clade; *E. huxleyi* occupied this place in the gene tree while two *Porphyra* strains replaced in the protein tree. The cryptophytes formed a separated clade in both protein and gene trees but not support.

Table 3.6.1: The distribution of *ycf*26 in photosynthetic plastomes

Taxon/organisms	Sensor-like histindine kinase accession number	Gene name	Protein name or product
Glaucophyte			
Cyanophora paradoxa	Not present		
Rhodophyte			
Porphyra purpurea	NP_054002	*ycf*26	hypothetical chloroplast ORF 26
Porphyra yezoensis	YP_537073	PoyeCp206	hypothetical chloroplast protein 26
Gracilaria tenuistipitata var.*liui*	YP_063707	*dfr*	drug sensory protein A
Cyanidium caldarium	Np_045190	*ycf*26	hypothetical protein
Cyanidioschyzon merolae	Not present		
Porphyridium aerugineum	unknown		
Rhodella violacea	unknown		
Bacillariophyte			
Odontella sinensis	Not present		
Phaeodactylum tricornutum	Not present		
Thalassiosira pseudonana	Not present		
Cryptophyte			
Guillardia theta	Not present		
Rhodomonas salina CCMP 1319	NC_009573	*ycf*26	two-component sensor kinase
C. erosa CCAC 0018	Newly sequenced		
C. bovoidea CCAC 0031	Newly sequenced		
Proteomonas sp. CCMP 0704	Newly sequenced		
Hemiselmis tepida CCMP 0443	Newly sequenced		
Teleaulax sp. SCCAP K-416	Newly sequenced		
Rhodomonas sp. M 1480	Newly sequenced		
Chroomonas sp. SAP 980-1	Newly sequenced		
Haptophyte			
Emiliania huxleyi	YP_277392	*dfr*	two component sensor kinase
Raphidophyte			
Heterosigma akashiwo CCMP 0452	YP_00193515	*tsg*1	His-Asp sensor kinase
Pelagophyte			
Aureoumbra laguensis	Not present		
Charophyte			
Chlorokybus atmophyticus	Not present		

Coppied and enriched from Duplessis et al. *(2007).*

Table 3.6.2: The positions and E-values of functional domains for cryptophytes *ycf*26 protein

	TMH1			PDS			TMH2			HAMP			PAS			HiKA			HATPase_C		
	Begin	**End**	**E-value**	**Begin**	**End**	**E-value**	**Begin**	**End**	**E-value**	**Begin**	**End**	**E-value**	**Begin**	**End**	**E-value**	**Begin**	**End**	**E-value**	**Begin**	**End**	**E-value**
SCCAP K-416																4	72	7.04e-02	119	234	2.4e+00
CCMP 0704																111	179	2.24e-07	226	347	8043e-12
CCAC 0018																200	269	2.72e-06	317	440	4.81e-07
CCAC 0031																102	172	3.85e-06			
CCMP 1319																128	196	1.7e-10	241	362	6.69e-15
M1480													5	17	3.09e+00	129	197	6.39e-12	242	317	2.43e+00
SAG 980-1																10	78	6.11e-09	123	241	7.76e-03
CCMP 0443																10	78	9.26e-09	123	239	2.96e-07
CCMP 0452																140	208	2.00e-13	253	369	1.85e-17
Emiliania huxleyi	26	48	-	51	174	***	178	200	-	177	246	9.06e-12	260	327	2.72e-03	386	454	4.95e-18	499	612	3.78e-23
Gracilaria tenuistipitata var. liui	22	44	-	47	185	***	189	211	-	188	257	8.90e-09	271	337	7.93e-05	397	465	1.03e-14	510	633	1.95e-25
Cyanidium caldarium	20	42	-	45	187	***	191	213	-	190	259	2.90e-11	273	339	1.90e-02	399	470	2.57e-16	515	631	2.10e-21
Porphyra purpurea	41	63	-	66	205	***	209	231	-	208	277	1.10e-15	291	357	2.85e-06	417	485	3.37e-20	530	654	3.80e-28
Porphyra yezoensis	38	60	-	63	202	***	2056	228	-	205	274	2.20e-15	288	354	1.45e-07	414	482	1.93e-20	527	651	2.47e-27
Synechocystis sp. PCC 6803	33	55	-	58	197	***	201	223	-	200	269	9.90e-16	283	349	1.47e-08-	422	490	5.4e-24	535	656	1.8e-36

(-) E-values are not supported on SMART server; (***) The positions were identified based on the codon alignment by SMART server and comparison with Synechocystis *sp. PCC 6803, thus E-values were not available; Empty boxes, unavailable sequences in the genome, thus SMART cannot identify.*

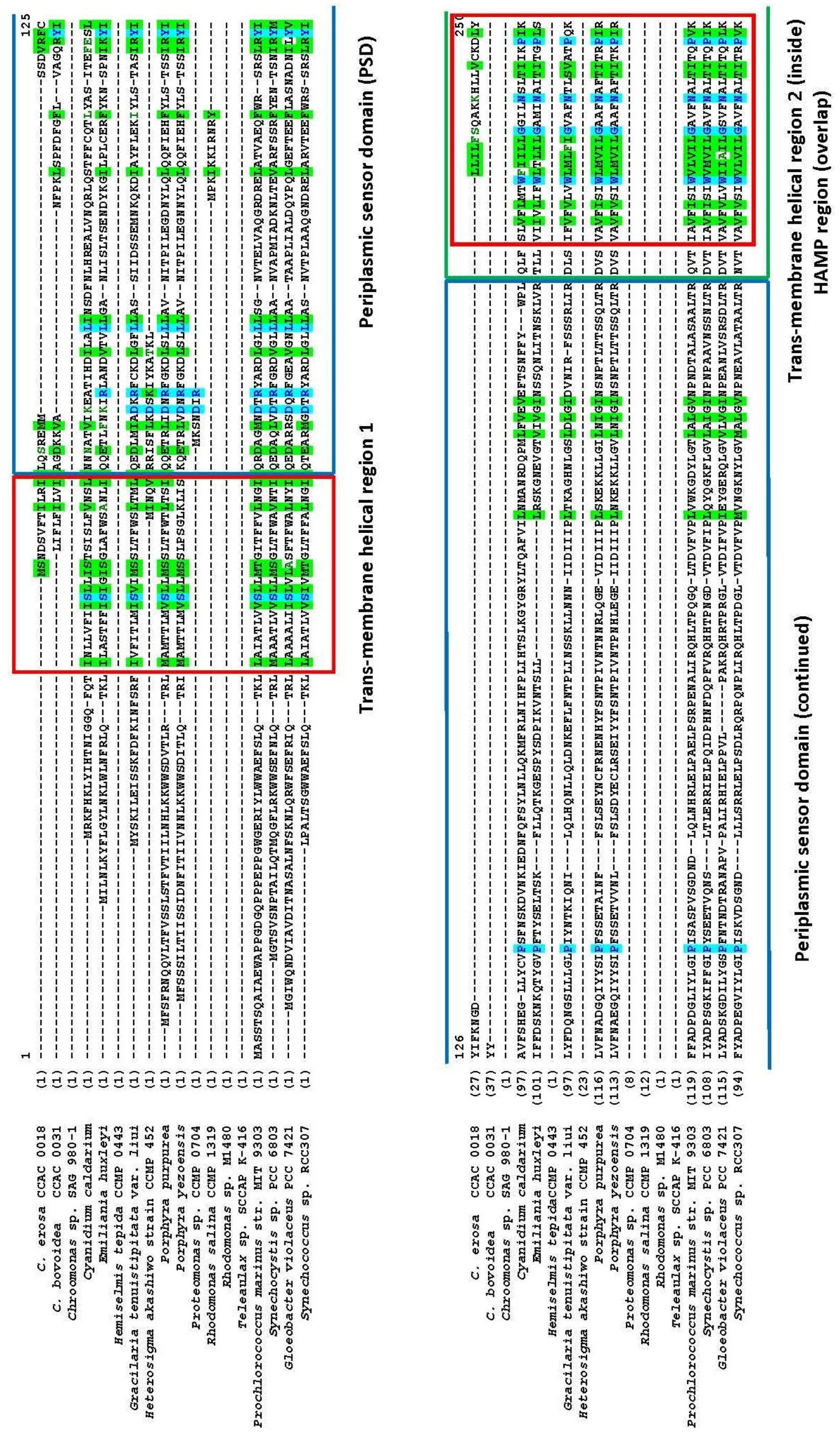

C. erosa CCAC 0018
C. bovoidea CCAC 0031
Chroomonas sp. SAG 980-1
Cyanidium caldarium
Emiliania huxleyi
Hemiselmis tepida CCMP 0443
Gracilaria tenuistipitata var. liui
Heterosigma akashiwo strain CCMP 452
Porphyra purpurea
Porphyra yezoensis
Proteomonas sp. CCMP 0704
Rhodomonas salina CCMP 1319
Rhodomonas sp. M1480
Teleaulax sp. SCCAP K-416
Prochlorococcus marinus str. MIT 9303
Synechocystis sp. PCC 6803
Gloeobacter violaceus PCC 7421
Synechococcus sp. RCC307
1
125
126
250
Trans-membrane helical region 1
Periplasmic sensor domain (PSD)
Periplasmic sensor domain (continued)
Trans-membrane helical region 2 (inside)
HAMP region (overlap)

Fig. 3.6.1: *Alignment for* ycf26 *proteins of cryptophytes and some random chosen Cyanobacteria.*

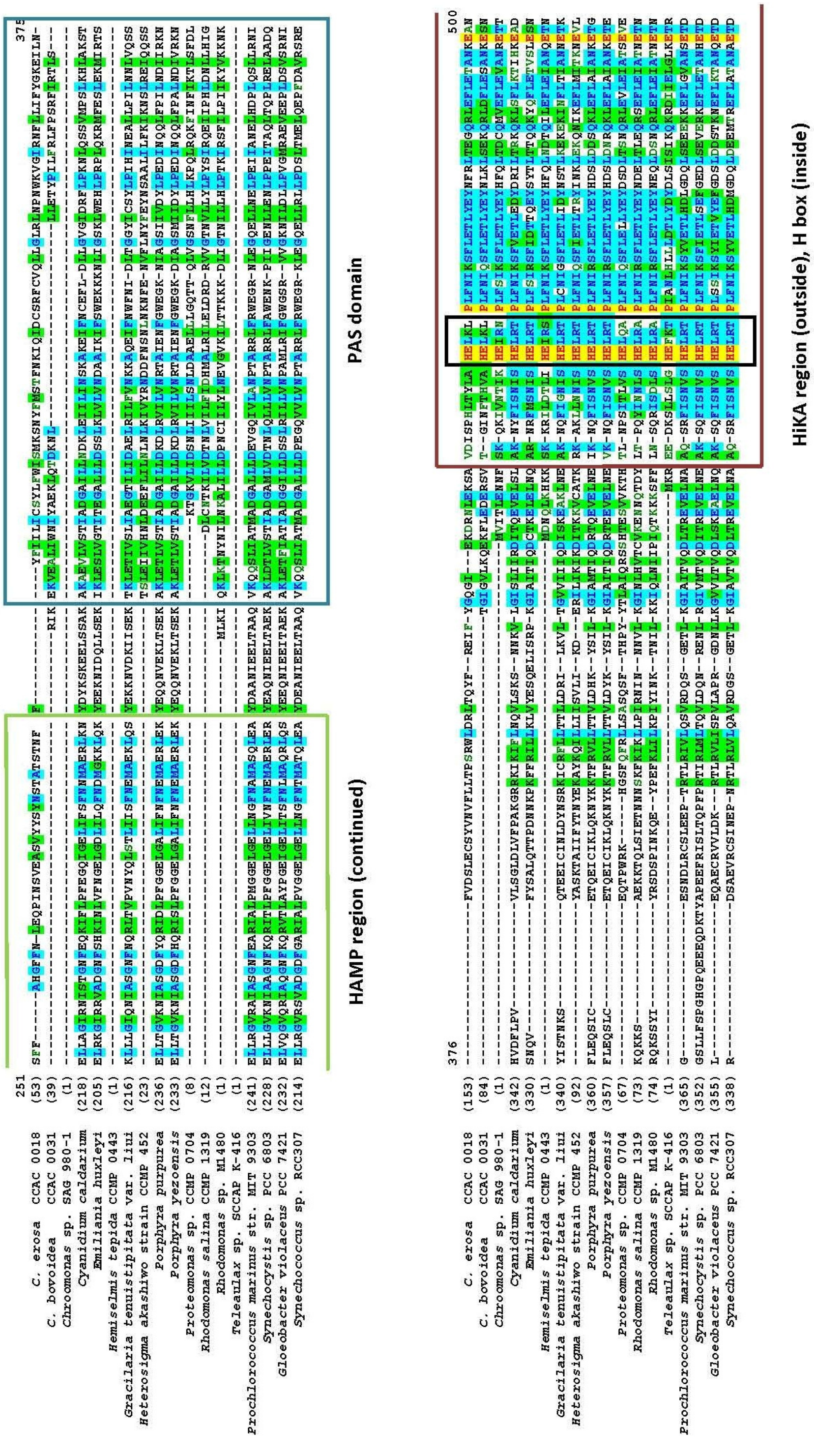


Fig. 3.6.1 *(continued): Alignment for ycf26 proteins of cryptophytes and some random chosen Cyanobacteria.*

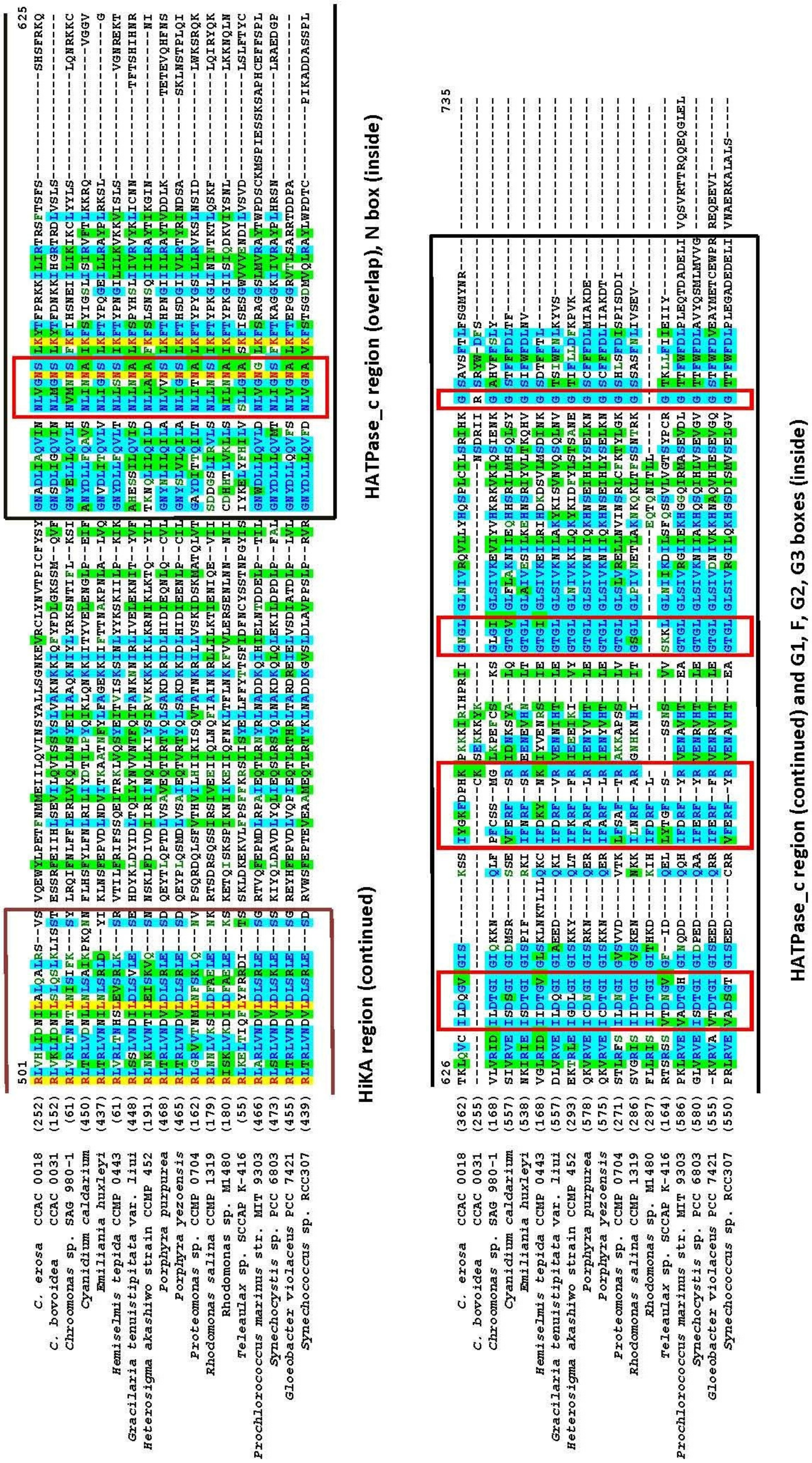

C. erosa CCAC 0018
C. bovoidea CCAC 0031
Chroomonas sp. SAG 980-1
Cyanidium caldarium
Emiliania huxleyi
Hemiselmis tepida CCMP 0443
Gracilaria tenuistipitata var. liui
Heterosigma akashiwo strain CCMP 452
Porphyra purpurea
Porphyra yezoensis
Proteomonas sp. CCMP 0704
Rhodomonas salina CCMP 1319
Rhodomonas sp. M1480
Teleaulax sp. SCCAP K-416
Prochlorococcus marinus str. MIT 9303
Synechocystis sp. PCC 6803
Gloeobacter violaceus PCC 7421
Synechococcus sp. RCC307
501
625
626
735
HiKA region (continued)
HATPase_c region (overlap), N box (inside)
HATPase_c region (continued) and G1, F, G2, G3 boxes (inside)

Fig. 3.6.1 *(continued): Alignment for ycf26 proteins of cryptophytes and some random chosen Cyanobacteria.*

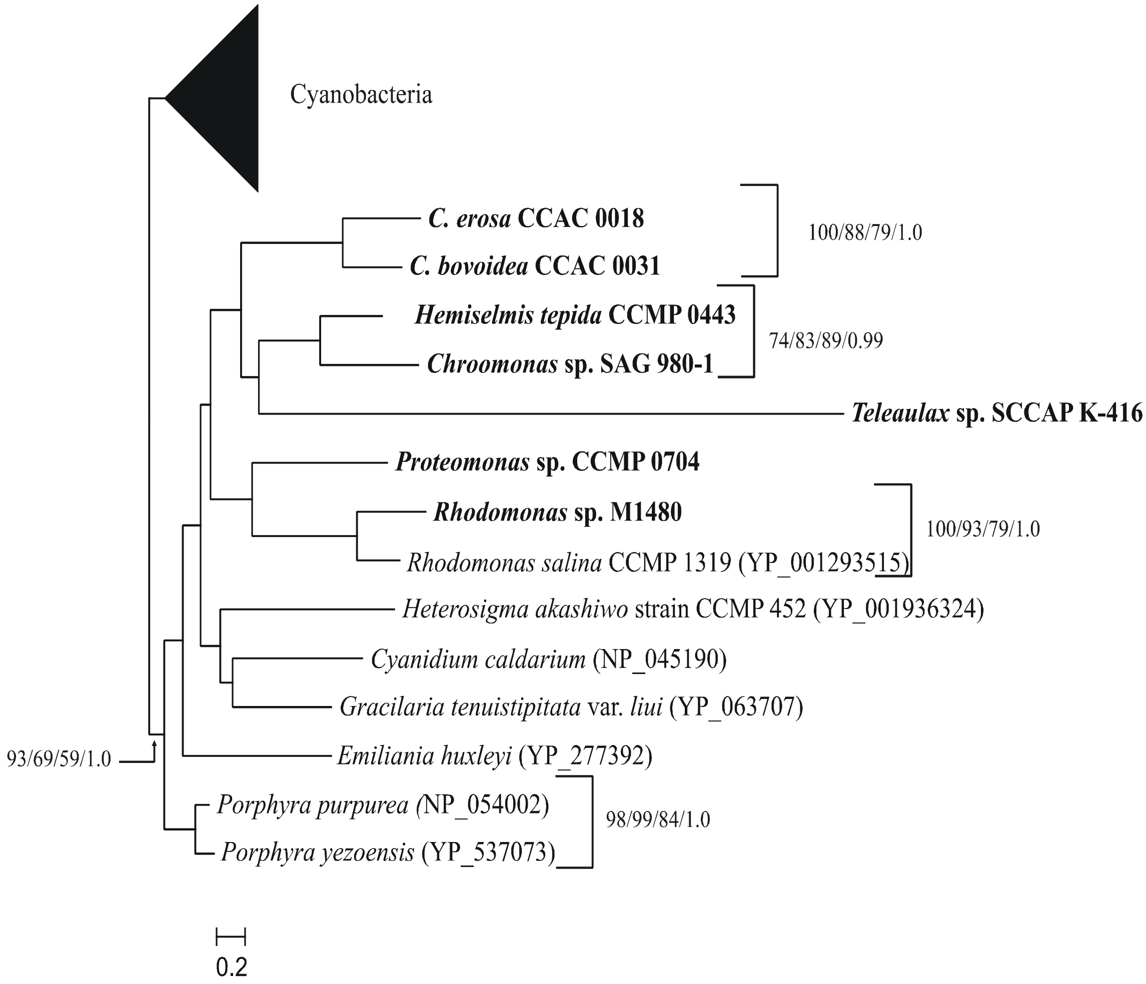

Cyanobacteria
C. erosa CCAC 0018
C. bovoidea CCAC 0031
100/88/79/1.0
Hemiselmis tepida CCMP 0443
Chroomonas sp. SAG 980-1
74/83/89/0.99
Teleaulax sp. SCCAP K-416
Proteomonas sp. CCMP 0704
Rhodomonas sp. M1480
Rhodomonas salina CCMP 1319 (YP_001293515)
100/93/79/1.0
Heterosigma akashiwo strain CCMP 452 (YP_001936324)
Cyanidium caldarium (NP_045190)
Gracilaria tenuistipitata var. *liui* (YP_063707)
93/69/59/1.0
Emiliania huxleyi (YP_277392)
Porphyra purpurea (NP_054002)
Porphyra yezoensis (YP_537073)
98/99/84/1.0
0.2

Fig. 3.6.2*: Rooted maximum-likelihood tree inferred from 59* ycf26 *protein sequences (154 characters). The evolutionary model (JTT+I+G, - ln L= 8192.52) were selected according to the results of the corrected Akaike Information Criterion (AICc) in ProtTest 1.4 without "+F" option. Bootstrap values were assigned from left to right: maximum likelihood/maximum parsimony/neighbor-joining/posterior probabilities. Names in bold faces: newly sequenced* ycf26 *in this study; scale bar = substitution per site.*

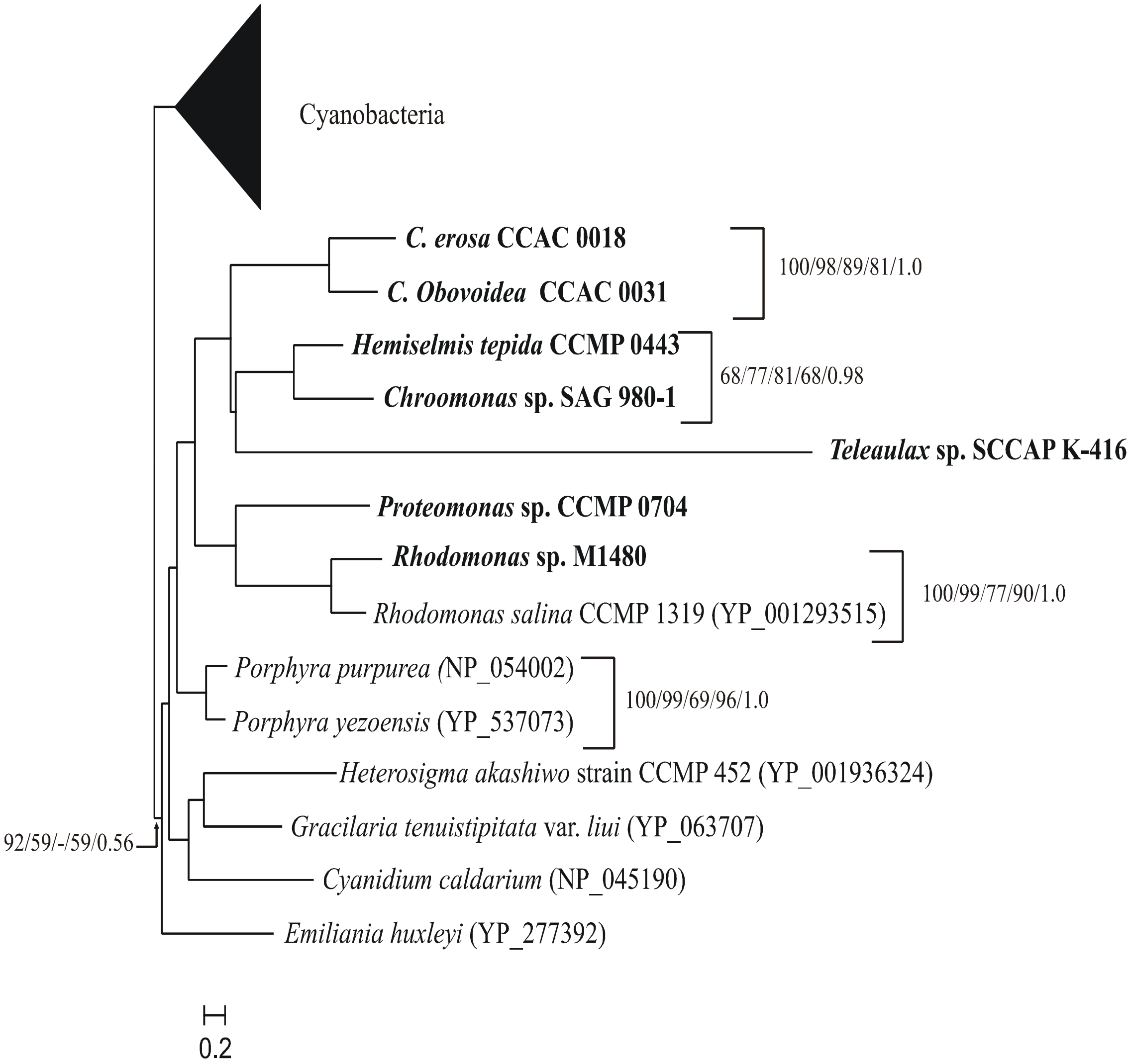

Cyanobacteria
C. erosa CCAC 0018
C. Obovoidea CCAC 0031
100/98/89/81/1.0
Hemiselmis tepida CCMP 0443
Chroomonas sp. SAG 980-1
68/77/81/68/0.98
Teleaulax sp. SCCAP K-416
Proteomonas sp. CCMP 0704
Rhodomonas sp. M1480
Rhodomonas salina CCMP 1319 (YP_001293515)
100/99/77/90/1.0
Porphyra purpurea (NP_054002)
Porphyra yezoensis (YP_537073)
100/99/69/96/1.0
Heterosigma akashiwo strain CCMP 452 (YP_001936324)
Gracilaria tenuistipitata var. liui (YP_063707)
92/59/-/59/0.56
Cyanidium caldarium (NP_045190)
Emiliania huxleyi (YP_277392)
0.2

Fig. 3.6.3: *Rooted maximum-likelihood tree reconstructed based on* ycf26 *gene sequences (59 taxa of 462 positions). The evolutionary model (TVM+I+G, - ln L= 17953.7167) were proposed by jModelTest 0.1 based on the corrected Akaike Information Criterion (AICc). From left to right: maximum likelihood bootstrap/maximum parsimony bootstrap/neighbor-joining bootstrap/logdet transformation bootstrap/posterior probabilities. Names in bold faces: newly sequenced* ycf26 *in this study; scale bar = substitution per site.*

3.7 ORF403

As described above, lying between *rps*4 and *rbc*L genes of Cryptophyta 16S rRNA–*rbc*L fragments were un-uniform sized ORFs. ORFs that occupy similar genomic location in the *G. theta* and/or *R. salina* plastomes are called ORF282 and/or ORF403 because putative proteins of these ORFs consisted of 282 or/and 403 amino acids, respectively (Douglas & Penny, 1999; Khan *et al.*, 2007a). Its structure and function in chloroplast of *Pisum sativum* var. Green Arrow were uncovered at first by Kouranov & Schnell (1997) and Kouranov *et al.* (1998). The authors described Tic22 protein as a 22-kD largely hydrophilic protein with no predicted transmembrane domains that was located in the intermembrane space between the outer and inner envelope membranes and was peripherally associated with the outer face of the inner membrane. The authors also proposed Tic22 acted as a receptor for precursor proteins as they emerged from the Toc complex (*t*ranslocon at the *o*uter envelope membrane of *c*hloroplasts), or mediate the association of Toc and Tic complexes (*t*ranslocon at the *i*nner envelope membrane of *c*hloroplasts) at contact sites. However, Tic22 has been less well defined in algae. In this part of study, Tic22 and ORF403 was used alternatively.

3.7.1 Verify the translated ORF403 of cryptophytes as Tic22 protein

All of Cryptophyta translated ORF403 nucleotides were put into SMART server to explore domain architectures and confidences. SMART recognized some Cryptophyta putative ORF403 proteins as Tic22 but very weak E-values (see Table 3.7.1). ORF403 of strains CCAC 0107, CCAC0109 CCAC0108, CCAC0113 and three colourless were not identified as Tic22 protein homologies. However, they were still kept in the ORF403 collection for some analyses in the next steps.

3.7.2 The distribution of Tic22 in host genomes

Using the Protein search function in NCBI database with Tic22 as the keyword, the finding showed interesting results of the distribution of the putative Tic22 proteins from Cyanobacteria to land-plants (Table 3.7.2).

At least one copy of the gene encoding for Tic22 was present in almost completed sequenced cyanobacterial genomes and plastomes of rhodophytes, cryptophytes such as those of *P. purpurea, P. yezoensis*, *G. tenuistipitata* var. *liui***,** *Cyanidium caldarium, Cyanidioschyzon merolae, G. theta, R. salina*, *Cryptomonas* sp. (this study), *etc*. One copy of this gene also was found in one haptophyte plastome, *E. huxleyi,* but nothing within chloroplast genomes of heterokont. The distribution of Tic22 was not limited in the plastomes; it also located in the other genetic machines such as nucleomorph or/and main nucleus. For example, the completely sequenced nucleomorph genomes of two cryptophyte *G. theta* and *H. andersenii* contain versions of Tic22 (Douglas *et al.*, 2001; Lane & Archibald, 2006a). In genome of *Cyanidioschyzon merolae*, the investigators identified the genes encode for Tic22 and Tic22-like proteins (CMJ105C and CMC181C) while an EST (expression sign tags) for Tic22 was found in the nucleus of *G. theta* (Matsuzaki *et al.*, 2004; Gould *et al.*, 2006). Interestingly, the Tic22 encoding gene was not located in the apicomplast but in nucleus of apicomplexan *Plasmodium falciparum (*Waller *et al.,* 2005*).*

While searching the documents for this part, an interesting paper published by Kalanon & McFadden in 2008 involving the chloroplast protein translocation complexes in plant, green algae and red algae were found According to this report, the copy of Tic22 encoding genes seemed to be more than one in land-plants. For example, the genomes of vascular plant *Arabidopsis thaliana* and non-vascular plant *Physcomitrella patents* had two Tic22 paralogs while other vascular plant *Pisum sativum* contained one Tic22 gene. Whereas, only one copy of this gene was found in the main nucleus of *Chlamydomonas reinhardtii* and no trace of Tic22 encoding gene was detected in other green algae *Ostreococcus lucimarinus* and *O. tauri*. The authors also argued the existence of Tic22 in the chloroplast genome of *Cyanidioschyzon merolae;* and did not include at least two copies of *Oryza sativa* Tic22 proteins into their database. Therefore, in this part of study, these abandon sequences were accepted for analyses. EST in genomes of some being sequenced organisms also verified the existences of Tic22. Hence, nuclear-encoded Tic22 version of non-photosynthetic green algal *Prototheca wickerhamii* was recognized (Borza *et al.*, 2005). Moreover, the EST database in NCBI supported that *Glycine max, Nicotina benthamiana, Avicennia marian, Gossypium hirsutum, Solanum tuberosum etc*. consisted of nuclear-encoded Tic22 genes; unfortunately, these EST coding for these Tic22 contained so many sequencing errors, so they were ignored in this analysis.

3.7.3 Reconstruct the phylogenetic trees

To take a closer look at the evolution of Tic22, attempts to reconstruct the phylogenetic trees based on the protein and nucleotide sequences were done. Most available Tic22 sequences in the NCBI data bank were downloaded and combined with newly sequenced Cryptophyta Tic22 to produce a large Tic22 data set of which 28 were Cyanobacteria, two were green algae, eight are land-plants, three were apicomplexan and 27 are red-line algae. Strains CCAC 0107 and CCAC0109 were not participated in this investigating because of their partial sequences.

The raw data set was submitted to Selecton server for codon-based alignment. Then, the alignable codon positions in both nucleotides and amino acid samples were extracted prior phylogenies. The nucleotide database consisted of 306 positions in length correspond with 102 amino acids. For nucleotide data set, jModeltest proposed GTR + G + I (General time reversible model with among-site substitution rate variation plus proportion of invariable sites) while ProtTest chose the CpREV + G (general time reversible model for plastid-encoded model gene) for protein data set.

Although the numbers of positions contained in data sets offered in phylogenetic analyses were not so long, the trees built on the protein and/or nucleotide database opened the first look about the evolution of Tic22 (Fig. 3.7.1 and 3.7.2).

In overall view, the sequences of Tic22 protein formed four main clades: the first clade was all Cyanobacteria, the second clade was those of red-line plastids, the third clade contained the Tic22 sequences that positioned in main nuclear genomes of green-algae (Chlorophyta) and land-plants (Magnoliophyta), the last clade comprised chromalveolate non-plastid Tic22 proteins (nucleomorph or/and nuclei genomes).

However, in the Tic22 gene tree, the branch pattern had some changes. The red-line and green-line Tic22 sequences were still present beside the Cyanobacteria clade, but the non-plastid clade was divided into two small groups: one contained all *Plasmodium* sequences and one was those of red-line non-plastid Tic22.

The bootstrap values for red-line Tic22 clade were obtained only in the protein but not in gene tree. In both trees, *C. borealis* CCAC0113, *C. erosa* CCAC 0108 and three *C. paramaecium* were always found in the *Cryptomonas* clade. The CCAC0113 displayed a significantly accelerated evolutionary rate than those of other *Cryptomonas* or even though of red-line ORF403. In the gene tree, three Tic22 of rhodophytes formed a clade at the basal position, but they were replaced by cluster of *E. huxleyi* and *Cyanidium caldarium* RK1 in the protein tree.

Like the red-line ORF430 sequences, the green-line Tic22 formed a separated clade with support values for both phylogenies trees (74%/57%/79%/0.84 and 94%/83%/-/-/0.99 for protein and nucleotide trees, respectively). Interestingly, non-photosynthetic *P. wickerhamii* and photosynthetic *C. reinhardtii* were always occupied at the basal positions in both phylogenetic trees.

The forming of chromalveote non-plastid genome Tic22 clade in the protein tree but not always support was an attractive result. This clade, however, was dispersed in the gene tree. The cluster contained nuclear *Cyanidioschyzon merolae*, nucleomorph *H. andersenii* nucleomorph *G. theta* and cluster contained three *Plasmodium* strains seemed to be relative while nuclear *G. theta* sequences was inserted into Cyanobacteria clade.

All Cyanobacterial Tic22 sequences grouped together as a super-clade but not always supported in the protein tree. They still formed a super clade in the gene tree but there were some strains jumped out of the Cyanobacterial cluster to insert into other clades. Surprisingly, the second version of *C. merolae* Tic22 (assigned Tic22-like protein - CMC181C) was always belong to Cyanobacteria clade in both phylogenetic trees.

3.7.4 Search for the conserved domains

The amino acid alignment showed that the alignable regions distributed mainly in the C-termini rather than in N-termini. These regions were considered to determine the conserved motifs. Unfortunately, the identification within these regions was not strict across all taxa. Thus, no region was assigned as the conserved motif. This contradicts to the report of Kalanon & McFadden (2008) that two conserved motifs were found in Tic22 proteins at 104 – 119 and 186 – 200 residues, respectively, according to *C. reinhardtii*. The explanation may be due to the limited numbers (nine sequences) of taxa distributed in nuclei genome in their survey versus the extended Tic22 collection of 70 sequences distributed in many kinds of genomes in this study. Incidentally, the annotation in the Figure 4 in their report had a minor mistake in that the sequences that were noted as PpTic22-1, PpTic22-2, AtTic22-III, TeTic22, CrTic22, CmTic22 and CmTic22-like have to chance to AtTic22-III, PpTic22-1, PpTic22-2, CrTic22, CmTic22.

3.7.5 Classify the Tic22 proteins

Base on the multi-alignment by Selecton server as well as the protein tree, the Tic22 protein sequences were divided into four groups.

The first group consisted of all Tic22 from Cyanobacteria. This group was characterized by the alignable N-terminus and polar-residue rich C-terminus. Exception for *Synechococcus* sp. PCC7002, all members of this class displayed a high homologous. The *Cyanidioschyzon merolae* Tic22-like protein was assigned into this group as its position the protein and gene trees was always in the Cyanobacteria clade.

The second group composed of Tic22 proteins located in non-plastid genomes of chromalveote (nucleomorph or/and nuclear genomes). Unlike group 1, the N-termini region of this group was high variant.

The third group contained Tic22 that positioned in nuclear genomes of green algae (Chlorophyta) and land-plants (Magnoliophyta). The non-photosynthetic *P.*

wickerhamii also belonged to this group. Like group 2, the N-termini region of group 3 was unalignable.

The fourth group included all Tic22 proteins originated from red-line plastids. This group was characterized by a specific conserved cluster containing 27 amino-acids and laying at about 125 residues from start Methionine, it was illustrated in Fig. 3.7.3 under the name "red-line plastid cluster". Moreover, exception for *G. theta*, *C. lundii* CCAC 0107, *C. borealis* CCAC 0113 and three colourless, the cryptophytes had a strictly identical block of about 25 amino-acids at the C-terminus – called "*Cryptophyte tail*".

The partial Tic22 sequences of CCAC 0107 and CCAC 0109 showed that their C-termini regions were alignable with other red-line Tic22 proteins. Especially, strain CCAC 0109 presented its *cryptophyte tail*. Strains CCAC 0113 and three colorless had no "*cryptophyte tail*" but possessed the red-line plastid cluster in their sequences. These verify that the putative translated ORF403 sequences of strains CCAC 0107, CCAC 0109, CCAC0113 and three colourless are actually Tic22 proteins.

As mentioned above, the translated sequence ORF403 of *Cryptomonas* colourless and CCAC 0113 were not recognized as the Tic22 protein by SMART server, the phylogenetic analyses in both protein and nucleotide data sets always showed that they actually belong to Cryptophyta clade with high-accelerated evolutionary rates. Moreover, three colorless always displayed a strong attachment to each other in a single group.

These result, again, confirmed that ORF403 of CAC 0107, CCAC0109 CCAC0108, CCAC0113 and colourless strains was obviously belong to ORF 403 family despite of the refutation of SMART server.

Table 3.7.1: E-values supported by SMART server for identify the Cryptophyta Tic22 proteins.

Strain	E values
CCA C0086	$6.00e^{-03}$
CCAC 0056	not recognized
CCAC 0018	$2.80e^{-02}$
CCAC 0107	not recognized
CCAC 0108	$2.10e^{-01}$
CCAC 0109	not recognized
CCAC 0113	not recognized
CCAC0031	$1.20e^{-04}$
CCAP 977/2a	not recognized
CCMP 0443	$1.90e^{-02}$
CCMP 0152	$7.70e^{-03}$
CCMP 0704	$4.70e^{-03}$
G. theta - nucleomorph	$5.40e^{-155}$
G. theta - plastid	$7.10e^{-02}$
H. andersenii - nucleomorph	$2.3e^{-26}$
M 1092	$4.50e^{-02}$
M 1480	$1.90e^{-04}$
M 2088	$6.60e^{-04}$
M 2452	not recognized
R. salina CCMP 1319	$8.60e^{-05}$
SAG 2013	$3.40e^{-04}$
SCCAP K416	$3.90e^{-04}$

Table 3.7.2: The number and location of Tic22 genes in algae and land plant genomes**.**

TAXON/ORGANISMS	NUCLEUS	PLASTID	NUCLEOMORPH
Glaucophyte			
Cyanophora paradoxa	Un clear	Not present	
Rhodophyte			
Porphyra purpurea	Un clear	NP_053828	
Porphyra yezoensis	Un clear	AP006715.1	
Gracilaria tenuistipitata var.*liui*	Un clear	YP_063677.1	
Cyanidium caldarium strain RK1	Un clear	NP_045131.1	
Cyanidioschyzon merolae	CMJ105C and CMC181C	AF02218	
Cryptophyte			
Guillardia theta	CAJ74146	NP_050704	AAK39834
Rhodomonas salina CCMP 1319	Un clear	YP_001293517	Un clear
Cryptomonas sp.	Un clear	***1 (newly sequenced)***	Un clear
Proteomonas CCMP 0704	Un clear	***1 (newly sequenced)***	Un clear
Hemiselmis tepida CCMP 0443	Un clear	***1 (newly sequenced)***	XP_001712614
Teleaulax SCCAP K-416	Un clear	***1 (newly sequenced)***	Un clear
Rhodomonas M 1480	Un clear	***1 (newly sequenced)***	Un clear
Chroomonas SAG 980-1	Un clear	***1 newly sequenced)***	Un clear
Haptophyte			
Emiliania huxleyi CCMP 373	Un clear	AAX13877	
Raphidophyte			
Heterosigma akashiwo CCMP 452	Un clear	Not present	
Bacillariophyte			
Odontella sinensis	Un clear	Not present	
Phaeodactylum tricornutum	Un clear	Not present	
Thalassiosira pseudonana	Un clear	Not present	
Apicomplexa			
Plasmodium falciparum 3D7	XP_001351847	Not present	
Chlorophyte			
Chlamydomonas reinhardtii	XP_001692709	Not present	
Prototheca wickerhamii	AY616047	Not present	
Prasinophyte			
Ostreococcus lucimarinus	Not present	Un clear	
Ostreococcus tauri	Not present	Not present	
Magnoliophyte			
Arabidopsis thaliana	NP_189013 and NP_195061	Not present	
Pisum savitum	AAC64606	Un clear	
Vitis vinifera	AM468354	Un clear	
Oryza sativa	NP_001059394 and BAD35192	Not present	
Bryophyte			
Physcomitrella patents	XM_001766060 and XM_001780538	Not present	

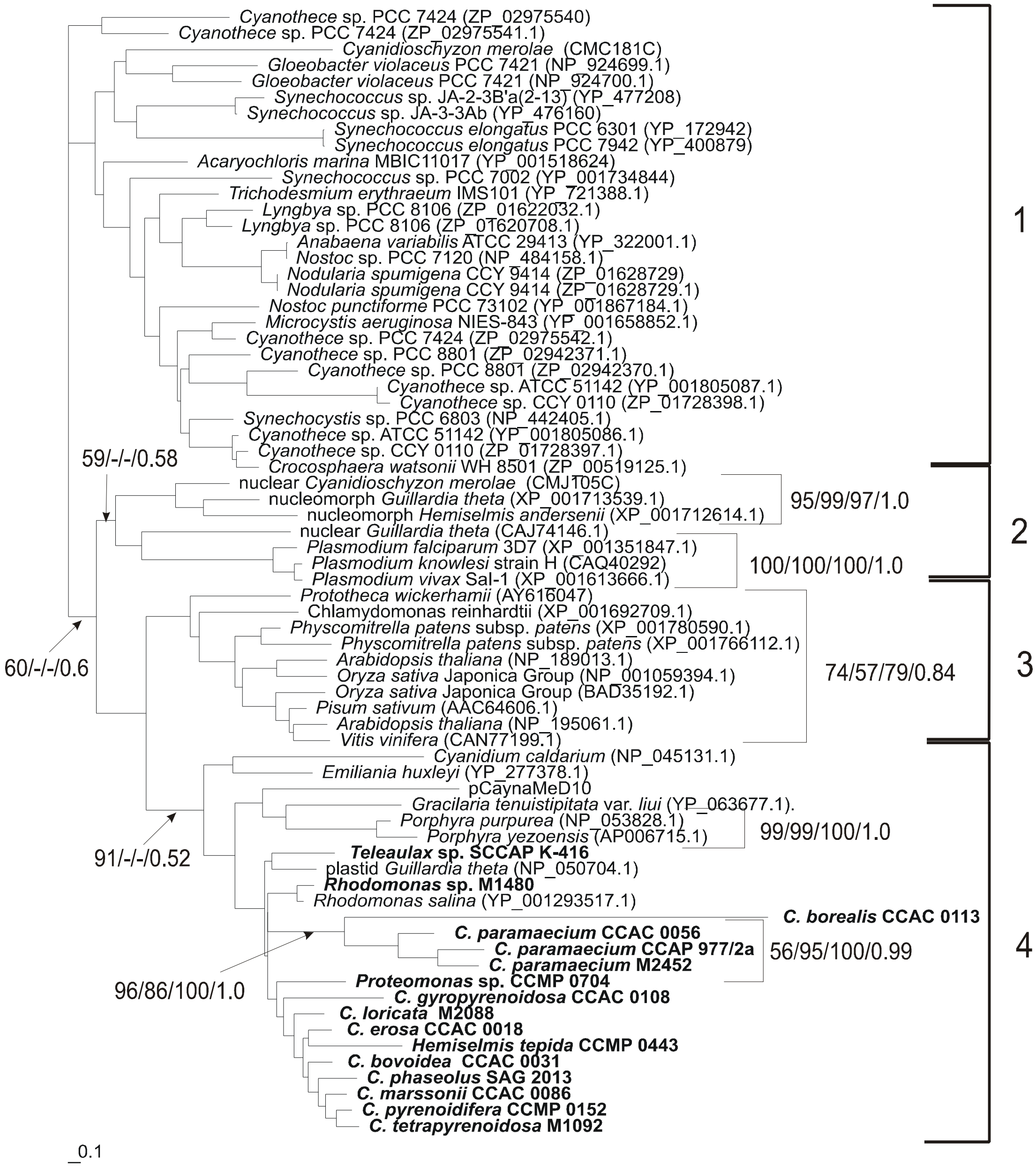

Cyanothece sp. PCC 7424 (ZP_02975540)
Cyanothece sp. PCC 7424 (ZP_02975541.1)
Cyanidioschyzon merolae (CMC181C)
Gloeobacter violaceus PCC 7421 (NP_924699.1)
Gloeobacter violaceus PCC 7421 (NP_924700.1)
Synechococcus sp. JA-2-3B'a(2-13) (YP_477208)
Synechococcus sp. JA-3-3Ab (YP_476160)
Synechococcus elongatus PCC 6301 (YP_172942)
Synechococcus elongatus PCC 7942 (YP_400879)
Acaryochloris marina MBIC11017 (YP_001518624)
Synechococcus sp. PCC 7002 (YP_001734844)
Trichodesmium erythraeum IMS101 (YP_721388.1)
Lyngbya sp. PCC 8106 (ZP_01622032.1)
Lyngbya sp. PCC 8106 (ZP_01620708.1)
Anabaena variabilis ATCC 29413 (YP_322001.1)
Nostoc sp. PCC 7120 (NP_484158.1)
Nodularia spumigena CCY 9414 (ZP_01628729)
Nodularia spumigena CCY 9414 (ZP_01628729.1)
Nostoc punctiforme PCC 73102 (YP_001867184.1)
Microcystis aeruginosa NIES-843 (YP_001658852.1)
Cyanothece sp. PCC 7424 (ZP_02975542.1)
Cyanothece sp. PCC 8801 (ZP_02942371.1)
Cyanothece sp. PCC 8801 (ZP_02942370.1)
Cyanothece sp. ATCC 51142 (YP_001805087.1)
Cyanothece sp. CCY 0110 (ZP_01728398.1)
Synechocystis sp. PCC 6803 (NP_442405.1)
Cyanothece sp. ATCC 51142 (YP_001805086.1)
Cyanothece sp. CCY 0110 (ZP_01728397.1)
Crocosphaera watsonii WH 8501 (ZP_00519125.1)
59/-/-/0.58
nuclear Cyanidioschyzon merolae (CMJ105C)
nucleomorph Guillardia theta (XP_001713539.1)
nucleomorph Hemiselmis andersenii (XP_001712614.1)
95/99/97/1.0
nuclear Guillardia theta (CAJ74146.1)
Plasmodium falciparum 3D7 (XP_001351847.1)
Plasmodium knowlesi strain H (CAQ40292)
Plasmodium vivax Sal-1 (XP_001613666.1)
100/100/100/1.0
Prototheca wickerhamii (AY616047)
Chlamydomonas reinhardtii (XP_001692709.1)
Physcomitrella patens subsp. patens (XP_001780590.1)
Physcomitrella patens subsp. patens (XP_001766112.1)
Arabidopsis thaliana (NP_189013.1)
Oryza sativa Japonica Group (NP_001059394.1)
Oryza sativa Japonica Group (BAD35192.1)
Pisum sativum (AAC64606.1)
Arabidopsis thaliana (NP_195061.1)
Vitis vinifera (CAN77199.1)
60/-/-/0.6
74/57/79/0.84
Cyanidium caldarium (NP_045131.1)
Emiliania huxleyi (YP_277378.1)
pCaynaMeD10
Gracilaria tenuistipitata var. liui (YP_063677.1).
Porphyra purpurea (NP_053828.1)
Porphyra yezoensis (AP006715.1)
99/99/100/1.0
91/-/-/0.52
Teleaulax sp. SCCAP K-416
plastid Guillardia theta (NP_050704.1)
Rhodomonas sp. M1480
Rhodomonas salina (YP_001293517.1)
C. borealis CCAC 0113
C. paramaecium CCAC 0056
C. paramaecium CCAP 977/2a
C. paramaecium M2452
56/95/100/0.99
96/86/100/1.0
Proteomonas sp. CCMP 0704
C. gyropyrenoidosa CCAC 0108
C. loricata M2088
C. erosa CCAC 0018
Hemiselmis tepida CCMP 0443
C. bovoidea CCAC 0031
C. phaseolus SAG 2013
C. marssonii CCAC 0086
C. pyrenoidifera CCMP 0152
C. tetrapyrenoidosa M1092
0.1
1
2
3
4

Fig. 3.7.1: *Rooted maximum-likelihood tree reconstructed based on 68 ORF403 protein sequences (102 characters). The evolutionary model (CpRev+G, - ln L= 11324.68) were selected according to the results of the corrected Akaike Information Criterion (AICc) in ProtTest 1.4 without "+F" option. Bootstrap values were assigned from left to right: maximum likelihood/maximum parsimony/neighbor-joining/posterior probabilities; names in bold faces: newly sequenced ORF403 in this study; scale bar = substitution per site; (1): cyanobacterial genomes; (2) chromalveote non-plastid genomes; (3) Chloroplast and Magnoliophyta nuclear genomes; (4)red-line plastomes.*

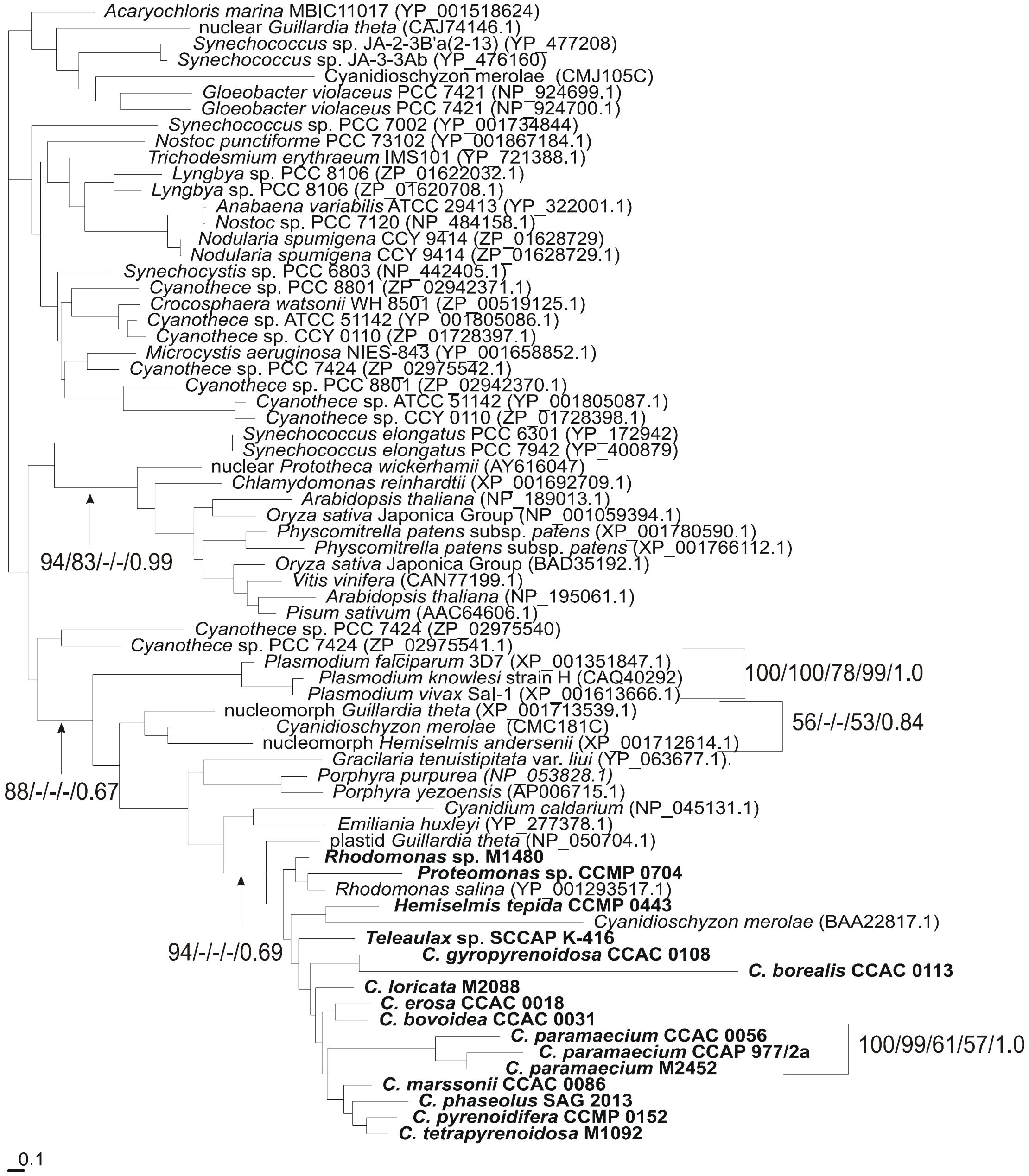

Acaryochloris marina MBIC11017 (YP_001518624)
nuclear Guillardia theta (CAJ74146.1)
Synechococcus sp. JA-2-3B'a(2-13) (YP_477208)
Synechococcus sp. JA-3-3Ab (YP_476160)
Cyanidioschyzon merolae (CMJ105C)
Gloeobacter violaceus PCC 7421 (NP_924699.1)
Gloeobacter violaceus PCC 7421 (NP_924700.1)
Synechococcus sp. PCC 7002 (YP_001734844)
Nostoc punctiforme PCC 73102 (YP_001867184.1)
Trichodesmium erythraeum IMS101 (YP_721388.1)
Lyngbya sp. PCC 8106 (ZP_01622032.1)
Lyngbya sp. PCC 8106 (ZP_01620708.1)
Anabaena variabilis ATCC 29413 (YP_322001.1)
Nostoc sp. PCC 7120 (NP_484158.1)
Nodularia spumigena CCY 9414 (ZP_01628729)
Nodularia spumigena CCY 9414 (ZP_01628729.1)
Synechocystis sp. PCC 6803 (NP_442405.1)
Cyanothece sp. PCC 8801 (ZP_02942371.1)
Crocosphaera watsonii WH 8501 (ZP_00519125.1)
Cyanothece sp. ATCC 51142 (YP_001805086.1)
Cyanothece sp. CCY 0110 (ZP_01728397.1)
Microcystis aeruginosa NIES-843 (YP_001658852.1)
Cyanothece sp. PCC 7424 (ZP_02975542.1)
Cyanothece sp. PCC 8801 (ZP_02942370.1)
Cyanothece sp. ATCC 51142 (YP_001805087.1)
Cyanothece sp. CCY 0110 (ZP_01728398.1)
Synechococcus elongatus PCC 6301 (YP_172942)
Synechococcus elongatus PCC 7942 (YP_400879)
nuclear Prototheca wickerhamii (AY616047)
Chlamydomonas reinhardtii (XP_001692709.1)
Arabidopsis thaliana (NP_189013.1)
Oryza sativa Japonica Group (NP_001059394.1)
Physcomitrella patens subsp. patens (XP_001780590.1)
Physcomitrella patens subsp. patens (XP_001766112.1)
94/83/-/-/0.99
Oryza sativa Japonica Group (BAD35192.1)
Vitis vinifera (CAN77199.1)
Arabidopsis thaliana (NP_195061.1)
Pisum sativum (AAC64606.1)
Cyanothece sp. PCC 7424 (ZP_02975540)
Cyanothece sp. PCC 7424 (ZP_02975541.1)
Plasmodium falciparum 3D7 (XP_001351847.1)
Plasmodium knowlesi strain H (CAQ40292)
Plasmodium vivax Sal-1 (XP_001613666.1)
100/100/78/99/1.0
nucleomorph Guillardia theta (XP_001713539.1)
Cyanidioschyzon merolae (CMC181C)
nucleomorph Hemiselmis andersenii (XP_001712614.1)
56/-/-/53/0.84
Gracilaria tenuistipitata var. liui (YP_063677.1).
Porphyra purpurea (NP_053828.1)
88/-/-/-/0.67
Porphyra yezoensis (AP006715.1)
Cyanidium caldarium (NP_045131.1)
Emiliania huxleyi (YP_277378.1)
plastid Guillardia theta (NP_050704.1)
Rhodomonas sp. M1480
Proteomonas sp. CCMP 0704
Rhodomonas salina (YP_001293517.1)
Hemiselmis tepida CCMP 0443
Cyanidioschyzon merolae (BAA22817.1)
Teleaulax sp. SCCAP K-416
94/-/-/-/0.69
C. gyropyrenoidosa CCAC 0108
C. borealis CCAC 0113
C. loricata M2088
C. erosa CCAC 0018
C. bovoidea CCAC 0031
C. paramaecium CCAC 0056
C. paramaecium CCAP 977/2a
100/99/61/57/1.0
C. paramaecium M2452
C. marssonii CCAC 0086
C. phaseolus SAG 2013
C. pyrenoidifera CCMP 0152
C. tetrapyrenoidosa M1092
0.1

Fig. 3.7.2: *Rooted maximum-likelihood tree inferred from ORF403 gene sequences (68 taxa of 306 positions). The evolutionary model (TVM+I+G, - ln L= 19876.3852) were proposed by jModelTest 0.1 based on the corrected Akaike Information Criterion (AICc). From left to right: maximum likelihood bootstrap/maximum parsimony bootstrap/neighbor-joining bootstrap/logdet transformation bootstrap/posterior probabilities. Names in bold faces: newly sequenced ORF403in this study; scale bar = substitution per site.*

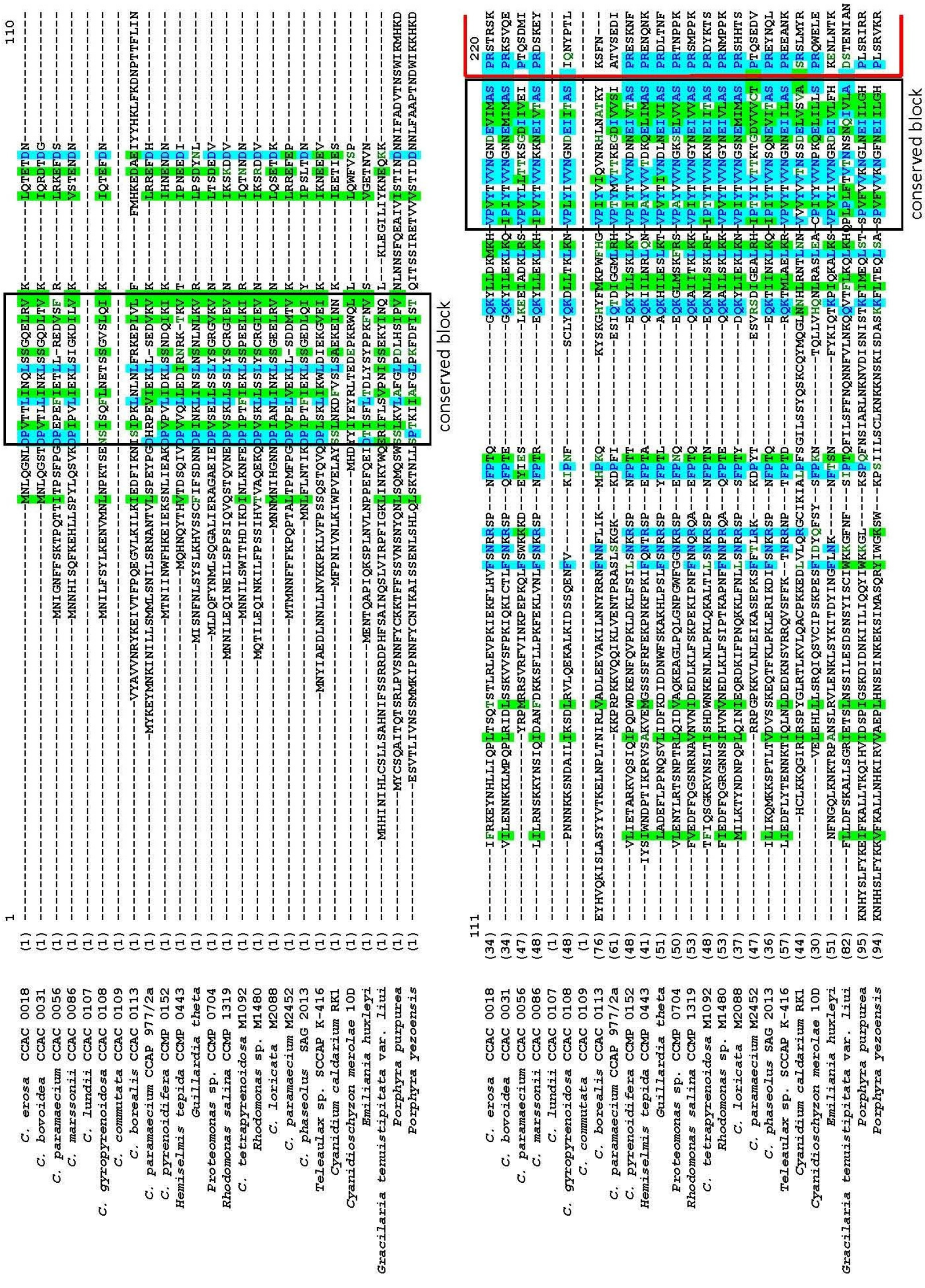


Fig. 3.7.3: *Alignment of Tic22 protein sequences originated from red-line plastids.*

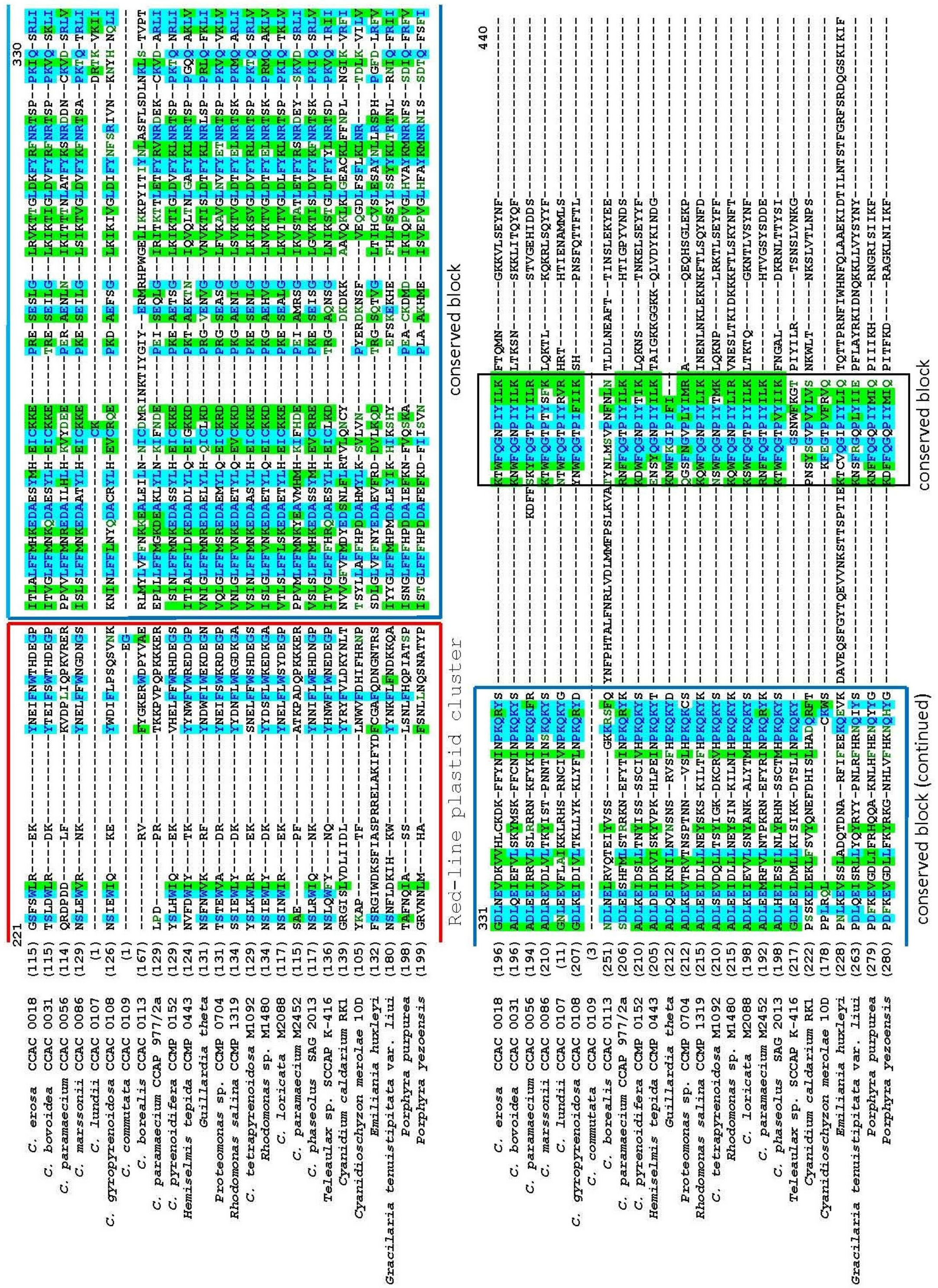


Fig. 3.7.3 *(continued): Alignment of Tic22 protein sequences originated from red-line plastids.*

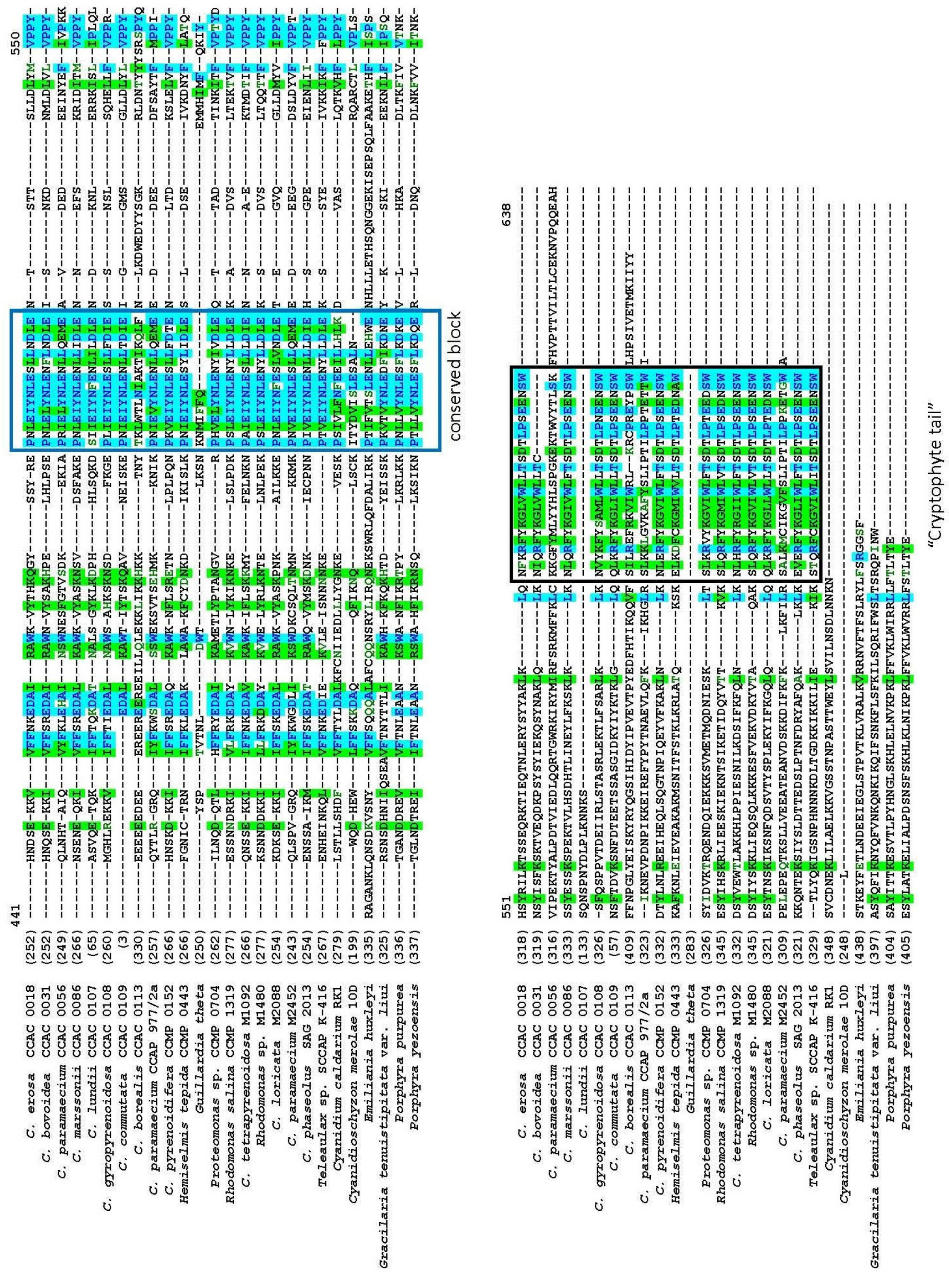


Fig. 3.7.3 *(continued): Alignment of Tic22 protein sequences originated from red-line plastids.*

3.8 Cluster of *psa*M and *chl*I gene and phylogenetic analyses based one the cryptophyte plastid *chl*I gene

3.8.6 **psa***M and* **chl***I as a conserved cluster in red-line plastid genomes*

Protein cluster analysis in NCBI server showed that *psa*M and *chl*I genes grouped together as a conserved cluster in some red-lineage algae such as *Odontella sinensis, Phaeodactylum tricornutum, Thalassiosira pseudonana* and *Porphyra purpurea, Porphyra yezoensis, Guillardia theta, Rhodomonas salina* and *Gracilaria tenuistipitata* var. *liui* (Fig. 3.8.1).

The absence of *psa*M genes from the fragments of 16S rRNA–*rbc*L of three colorless strains (M2452, CCAC 0056 and CCAP 977/2a) examined in this study is unexpected. Due to the lack of nucleotide information of the 16S rRNA–*chl*I fragments of colorless strain M1634, the presence or absence of *psa*M from M1634 is therefore questionable.

3.8.7 Annotate the conserved domains of **chl***I protein*

The structural and functional information of *psa*M in algae chloroplast have not been documented. Therefore, only the conservation of *chl*I protein was considered by observation the functional and conserved domains. To do that, the previously produced data sets were added by 11 red-lineage algae and one Cyanobacterium; the M1634 *chl*I, though was sequenced one third at C terminal only, was also added to have the preliminary view of its structure in compared with other cryptophytes.

The codon-alignment results showed all *chl*I predicted protein sequences contained four highly conserved regions at the positions of 11 – 57, 106 – 212, 280 – 306 and 318 – 338 as described for *chl*I of *C. phi* (previous name of *G. theta*) (Nakajama *et al.*, 1995). The conservation of these regions was very strict in photosynthetic algae but relaxed in colorless strains. Moreover, the divergence of *chl*I protein sequences in heterotrophic *Cryptomonas* seemed to be very great in the last 140 amino acids.

The first conserved region contained an ATP/GTP-binding site (GDRGTGKST) located at 45 – 53 positions of *G. theta chl*I. The second region also held other ATP/GTP-biding motif (ILYVDEVN) in between of 147 – 154 positions of *G. theta chl*I. Unsurprisingly, all plastid *chl*I proteins displayed a significant conservation in both ATP/GTP-binding motifs in which the second motif was completely conserved in all *chl*I proteins and the first motif had several minor changes in those of three colorless strains. Strain M 1634 was not read for the N-termini part of its *chl*I gene, so the first ATP/GTP-binding motif has to be waited until this *chl*I gene was read fully.

The codon-alignment also displayed two variant regions in all of *chl*I protein sequences that located between the first and second and the second and third conserved regions, respectively. The first variant region was equivalent to the insertion region 1, while the second one was correspond to helices *6 and *7 (see Fodje *et al.,* 2001 for more detail about three-dimension structure of *chl*I protein). In this study, a deletion of around 7 – 10 amino acid residues in the first variant region was found in colorless strains CCAC 0056 and CCAP 977/2a, respectively. Other point, the *chl*I protein of photosynthetic *H. akashiwo* CCMP 0452 also was suffered a cluster of 14 amino acids in this region. All four leukoplast-bearing *Cryptomonas* strains showed a significantly alteration in the second variant region.

Valentine *et al.* (1998) analyzed the N-termini of *chl*I protein to assume the localization of this protein in the cell. According to their information, the signal sequences were presented in the N-termini of *chl*I proteins. The first hydrophilic region with variant length contained basic residue such as K and R, the second region was hydrophobic with 10 – 11 amino acid residues, and the third hydrophilic region comprised five residues. The protein alignment result showed that the loss of 15 – 17 amino acid residues in two colourless strain M2452 and CCAP 977/2a were correspond to the hydrophilic region.

The relaxation in the last 140 amino acid residues and the divergence in the second variant region seemed to be the main factor caused the reducing of the identity degree of colorless *chl*I in comparison with those of photosynthetic cryptophytes

3.8.8 Reconstruct phylogenetic trees

To build in the phylogenetic trees based on the *chl*I sequences, the partial sequence of colorless M1634 was rejected while *P. purpurea*, *P. yezonensis* and *G. tenuistipitata* var.

liui were chosen as the out-group. The alignment, therefore, contained 24 taxa of 963 positions. The constant, variable and parsimony informative were counted and showed in the Table 3.8.1.

The number of variable sites at the third codon position was most extensive, nearly 100%; however, dropped down to 67.3% and 44.4% for the first and second codon positions, respectively. Moreover, almost all variation at third codon positions was phylogenetically informative (98.7% of variable sites) while the parsimony informative sites at the first and second codon positions occurred at 63.8% and 78.76%, respectively. These suggested that many of the sense changes were concentrated at the extreme terminal node of the tree(s). The high number of A and T (TA % = 77.7) at the third codon position indicated that a marked bias towards codons ending in TA.

The pruned data set with only cryptophytes showed that the number of variable and parsimony informative site nearly the same with the case of full data set.

When all taxa were included, the chi-square test of homogeneity of base frequencies results in significant P-values (Chi-square = 99.293971, df=69, P = 0.00988196) until CCAP 977/2a was deleted from the data set (Chi-square = 58.192941, df=51, P = 0.22766298).

The full data set was retested for the homogeneity with the first, the second position and the first two-codon positions (Chi-square = 57.116624, df=69, P = 0.84581378; Chi-square = 9.423643, df=69, P = 1.00000000 and Chi-square = 37.471928, df=69, P = 0.99930335, respectively).

To test what sequence(s) in the cryptophytes caused the bias, the collection of only cryptophytes was introduced to PAUP* to calculate the P-value again. As expected, the full cryptophyte data failed the test (Chi-square = 86.584238, df=60, P = 0.01395624). The extreme sequences (maxima and minima) in the T nucleotide content were removed one by one before searching the P-values. The collection still failed the test ($P<0.05$) when CCAC0109 , CCAC0064, SCCAP K-416, CCMP 0443 and *G. theta* were deleted but the P-value increased to 0.13014625 when colorless CCAC0056 excluded (Chi-square = 52.417327, df=42), and it jumped to 0.69342256 when the second colorless strain CCAP977/2a was removed from the data set (Chi-square = 34.083432, df=39). Exclusion of the last colorless strains M2452 lead the slightly increased in the P-values (Chi-square = 28.270793, df=36, P = 0.81741040). That suggested the colourless strains were the main factor affecting the homogeneity of base composition in the cryptophyte clade.

The jModeltest 0.1.1 proposed GTR+I+G as the best-fitting evolutionary model for *chl*I gene data set; and accompanied relative parameters were listed below.

Model = GTR+I+G
partition = 012345
-lnL = 14241.4952
K = 56
freqA = 0.4091
freqC = 0.1250
freqG = 0.1359
freqT = 0.3301
R(a) [AC] = 6.1772
R(b) [AG] = 8.5566
R(c) [AT] = 0.0167
R(d) [CG] = 3.2582
R(e) [CT] = 21.6152
R(f) [GT] = 1.0000
p-inv = 0.0460
gamma shape = 0.2570

Meanwhile, the amino acid data set was imported to ProtTest 1.4 to search for the suitable model without F option. The CpREV+G model was proposed for *chl*I protein data set.

The maximum likelihood tree of *chl*I gene obtained by PhyML revealed some un-expected points. The first point was the very high-accelerated evolutionary rate of colorless strains – M2452, CCAC 0056 and CCAP 977/2a – in comparison with other taxa in entire data set. The second point was that they occurred at the basal position in the *Cryptomonas* clade without close relative.

Unfortunately, the strain CCAC0107 and CCAC0108 were not presented in the *chl*I data set, their positions in the evolutionary tree, therefore, were not able to be determined. However, the current position of CCAC0113 and colourless strains in the *chl*I gene tree suggested that the LB clade found in *rbc*L gene tree seemed to be dispersed strongly.

The branching for *Cryptomonas* in the *chl*I gene tree was incongruent with those of *rbc*L gene tree (Hoef-Emden *et al.*, 2005) and concatenated nuclear partial LSU rDNA and nucleomorph SSU rDNA gene tree (Hoef-Emden, 2008). Exception for colorless clade with very high support values, there were only several clusters obtained the support values such as CCAC0006-M1092 (77%/63%/63%/79%/0.66, Fig. 3.8.3), CCAC0109-M2088 (94%/76%/79%/73%/1.0) and CCAC0031- CCAC0109-M2088 (71%/57%/59%/55%/0.98).

The branching of *chl*I protein tree, otherwise, was conflict with the gene tree (Fig. 3.8.4). The basal position in the *Cryptomonas* clade belonged to CCACC0113 while the three colorless strains moved inside grouping with clade M2088-CCAC0109 (76%/60%/57%/0.87) to form a moderate clade without support.

3.8.9 Codon usage

Observation the codon usage of amino acids coded by NNY codon (two-fold degenerate codon) in *chl*I gene found that NNU codon was preferred over codon NNC in all cryptophytes. Also, there was a shift in codon usage from RNN to YNN codons in three heterotrophic strains. In autotrophic cryptophytes, the numbers of RNN codon were always around 99 – 108 and those of YNN codon were around 213 – 222 while these numbers significantly increased or/and decreased in three colourless strains (123 – 133 and 188 – 198 for RNN and YNN codon, respectively).

Table 3.8.1: Base composition and variability of *chl*I sequences

	ALL TAXA (24 TAXA)								CRYPTOPHYTES (21 TAXA)			
	T%	C%	A%	G%	AT%	Conserved sites	Variable sites	Parsimony-Informative sites	Conserved sites	Variable sites	Parsimony-Informative sites	Total
1st codon position	16.4	16.5	30.8	36.3	47.2	105	216	170 (78.7% of variable sites)	110	211	166 (78.7 %)	321
2nd codon position	31.5	20.3	32.5	15.7	64	188	133	85 (63.9 %)	189	132	76 (57.6 %)	321
3rd codon position	37.9	10.0	39.8	12.3	77.7	2	319	313 (98.2 %)	4	317	309 (97.5 %)	321
All sites	28.6	15.6	34.4	21.4	63.0	295	668	568 (84.7 %)	303	660	551 (83.5 %)	963

Table 3.8.2: Codon usage of RNN, YNN and NNY codon in cryptophytes *chl*I gene**.**

	STRAIN	YNN	RNN	UAC/UAU (Tyr)	CAC/CAU (His)	AAC/AAU (Asn)	GAC/GAU (Asp)	UGC/UGU (Cys)	UUC/UUU (Phe)
01	M2452	124	197	4/4	3/7	6/14	7/9	4/4	3/9
02	CCAC 0056	133	188	2/5	2/7	6/11	3/15	2/5	3/7
03	CCAP9 77/2a	123	198	2/1	5/5	5/8	5/13	5/3	1/11
04	CCAC 0113	108	213	2/2	1/5	2/9	8/22	0/4	1/10
05	CCAC 0018	101	212	**3/2**	1/4	5/6	5/20	0/3	2/8
06	CCAC 0031	103	218	1/4	**4/1**	2/9	6/21	0/3	0/10
07	M2088	102	219	**3/2**	2/3	5/4	9/15	0/3	0/11
08	CCAC 0109	104	217	**3/2**	2/3	5/5	9/16	**2/1**	5/5
09	SAG 2013	102	219	0/5	1/4	4/7	3/24	1/2	0/10
0	CCAC 0086	98	223	**4/1**	1/4	6/7	7/18	**2/1**	1/9
11	CCAC 0006	108	213	1/4	1/4	3/10	4/21	0/3	1/9
12	CCAC 0064	97	224	1/4	0/5	2/10	2/27	0/3	2/9
13	M1092	103	218	1/4	1/4	2/9	7/19	0/3	3/7
14	CCMP 1052	101	220	2/3	1/4	4/7	4/20	1/2	2/8
15	SCCAP K-416	99	222	2/3	0/5	**5/4**	5/22	2/2	2/9
16	CCMP 0443	103	218	**3/2**	0/4	4/6	8/22	1/3	1/9
17	CCMP 0704	110	211	**3/1**	**3/2**	4/5	11/17	1/3	4/8
18	CCMP 1319	99	222	2/3	2/4	5/22	4/22	0/4	3/7
19	M1480	100	221	2/3	2/3	5/10	3/22	1/3	3/7
20	SAG 980-1	103	218	1/3	0/7	3/8	11/19	2/2	1/10

The chl*I data set used in phylogenetic analysis was used for counting RNN, YNN and NNY codon per 321 codon (963 nucleotide positions); bold face values indicate some exception in which NNC over NNT codon.*

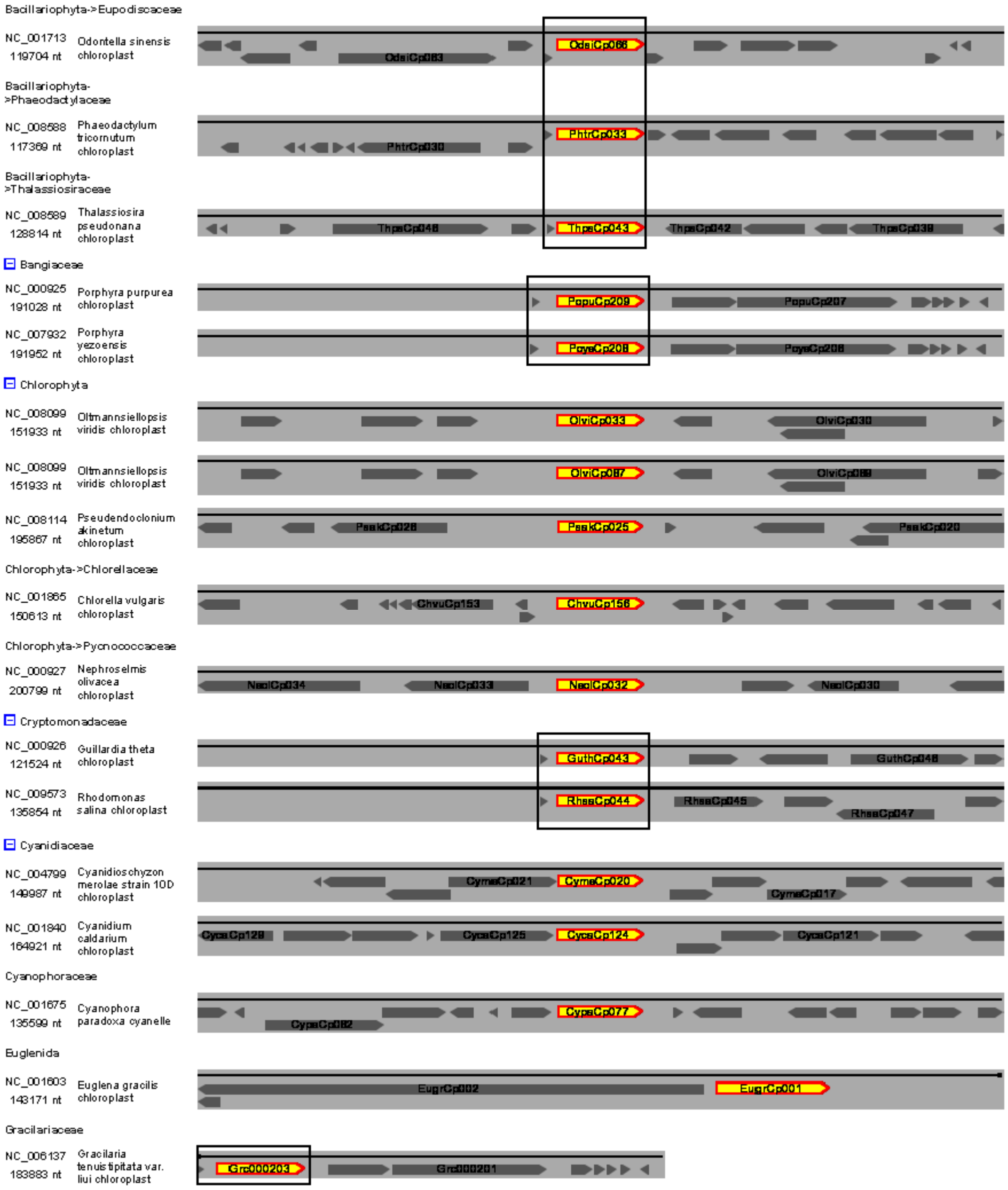

Bacillariophyta->Eupodiscaceae
NC_001713 119704 nt Odontella sinensis chloroplast
OdsiCp083
OdsiCp086
Bacillariophyta->Phaeodactylaceae
NC_008588 117369 nt Phaeodactylum tricornutum chloroplast
PhtrCp030
PhtrCp033
Bacillariophyta->Thalassiosiraceae
NC_008589 128814 nt Thalassiosira pseudonana chloroplast
ThpsCp046
ThpsCp043
ThpsCp042
ThpsCp039
Bangiaceae
NC_000925 191028 nt Porphyra purpurea chloroplast
PopuCp209
PopuCp207
NC_007932 191952 nt Porphyra yezoensis chloroplast
PoyeCp208
PoyeCp206
Chlorophyta
NC_008099 151933 nt Oltmannsiellopsis viridis chloroplast
OlviCp033
OlviCp030
NC_008099 151933 nt Oltmannsiellopsis viridis chloroplast
OlviCp087
OlviCp089
NC_008114 195867 nt Pseudendoclonium akinetum chloroplast
PsakCp028
PsakCp025
PsakCp020
Chlorophyta->Chlorellaceae
NC_001865 150613 nt Chlorella vulgaris chloroplast
ChvuCp153
ChvuCp156
Chlorophyta->Pycnococcaceae
NC_000927 200799 nt Nephroselmis olivacea chloroplast
NeolCp034
NeolCp033
NeolCp032
NeolCp030
Cryptomonadaceae
NC_000926 121524 nt Guillardia theta chloroplast
GuthCp043
GuthCp048
NC_009573 135854 nt Rhodomonas salina chloroplast
RhsaCp044
RhsaCp045
RhsaCp047
Cyanidiaceae
NC_004799 149987 nt Cyanidioschyzon merolae strain 10D chloroplast
CymeCp021
CymeCp020
CymeCp017
NC_001840 164921 nt Cyanidium caldarium chloroplast
CycaCp128
CycaCp125
CycaCp124
CycaCp121
Cyanophoraceae
NC_001675 135599 nt Cyanophora paradoxa cyanelle
CypaCp082
CypaCp077
Euglenida
NC_001603 143171 nt Euglena gracilis chloroplast
EugrCp002
EugrCp001
Gracilariaceae
NC_006137 183883 nt Gracilaria tenuistipitata var. liui chloroplast
Grc000203
Grc000201

Fig. 3.8.1: *The Genome ProtMap (NCBI) searching results showed the* psa*M*-chl*I clusters in some red-lineage algae such as* Odontella sinensis, Phaeodactylum tricornutum, Thalassiosira pseudonana *and* Porphyra purpurea, Porphyra yezoensis, Guillardia theta, Rhodomonas salina *and* Gracilaria tenuistipitata *var.* liui. *The light arrows in the boxes are* chl*I genes and the gray smaller arrows just behind the* chl*I are* psa*M genes.*

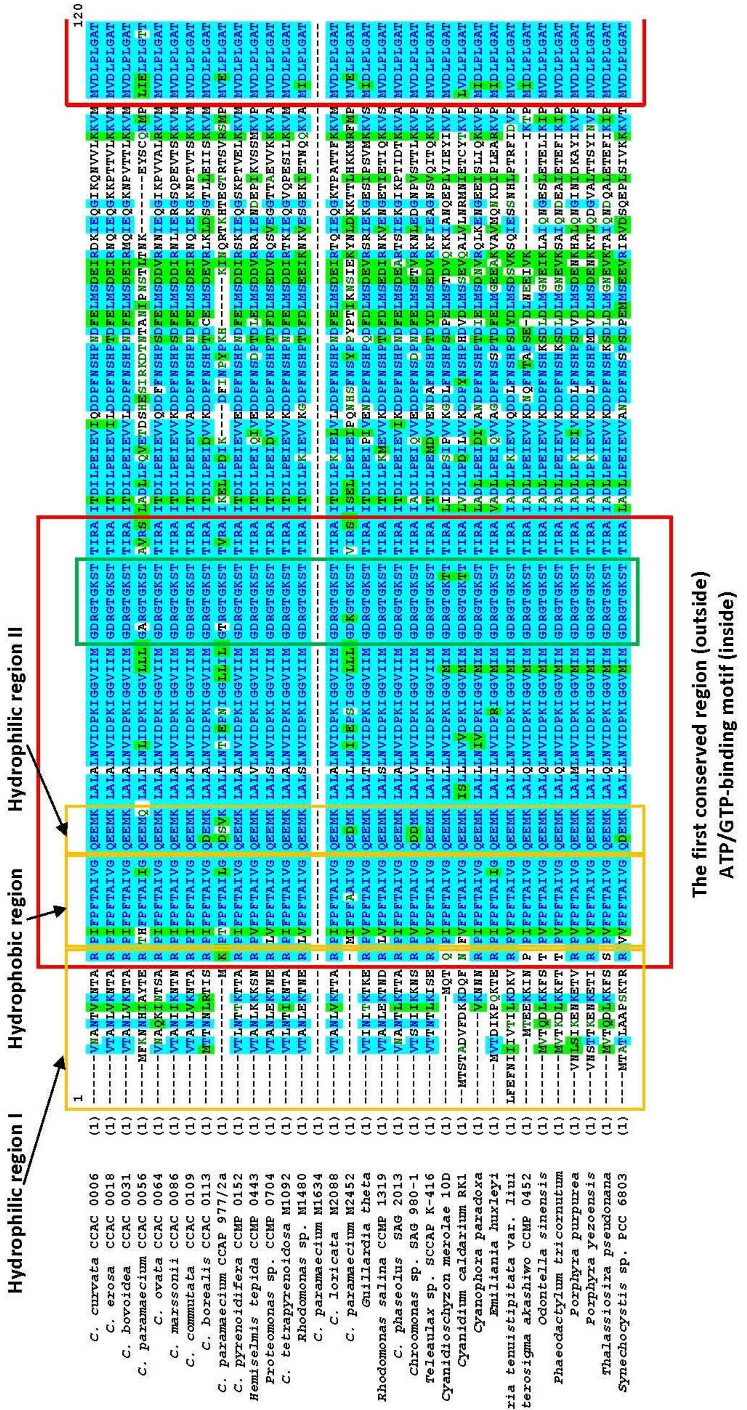


Fig. 3.8.2: *The annotating for the conserved regions and hydrophilic and hydrophobic regions of* chl*I protein were based on codon-alignment of sequences from cryptophytes, Cyanobacteria and some red-line plastids.*

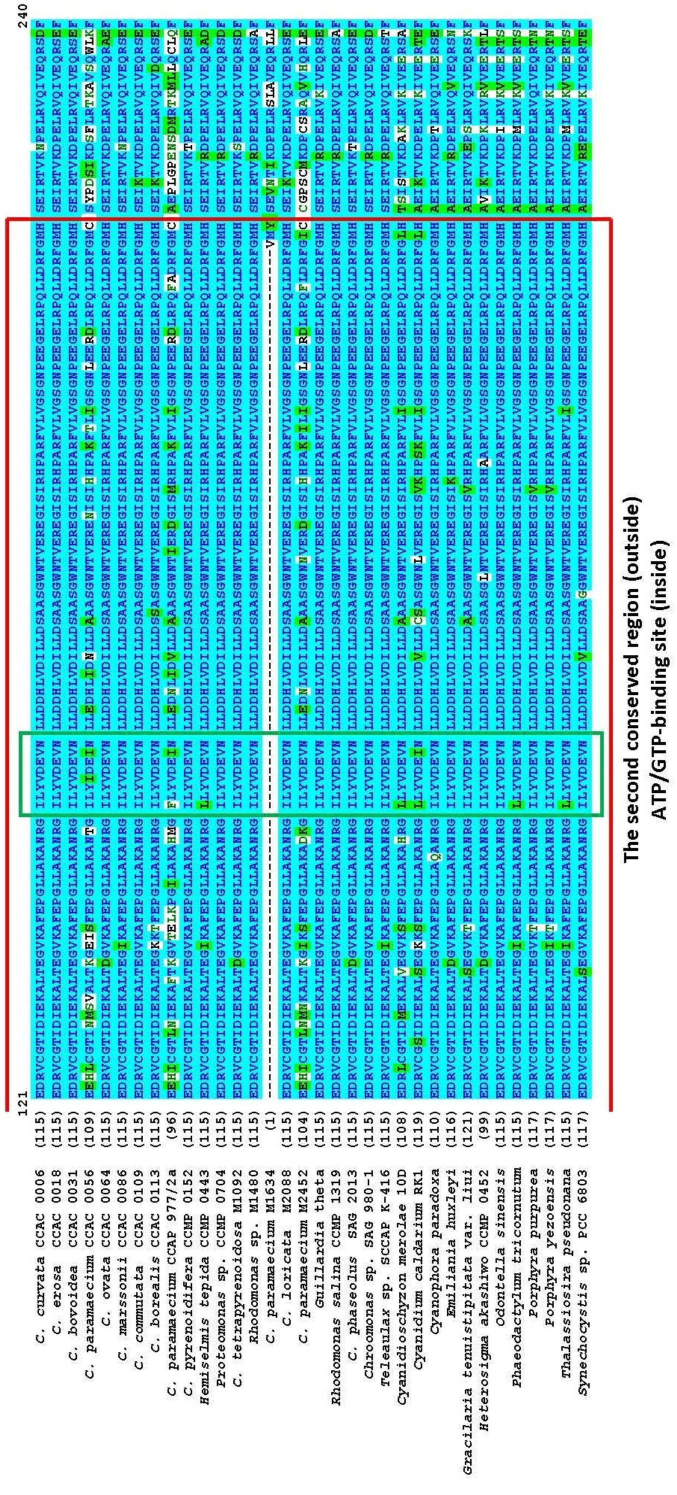

121
240
C. curvata CCAC 0006 (115)
C. erosa CCAC 0018 (115)
C. bovoidea CCAC 0031 (115)
C. paramaecium CCAC 0056 (109)
C. ovata CCAC 0064 (115)
C. marssonii CCAC 0086 (115)
C. commutata CCAC 0109 (115)
C. borealis CCAC 0113 (115)
C. paramaecium CCAP 977/2a (96)
C. pyrenoidifera CCMP 0152 (115)
Hemiselmis tepida CCMP 0443 (115)
Proteomonas sp. CCMP 0704 (115)
C. tetrapyrenoidosa M1092 (115)
Rhodomonas sp. M1480 (115)
C. paramaecium M1634 (1)
C. loricata M2088 (115)
C. paramaecium M2452 (104)
Guillardia theta (115)
Rhodomonas salina CCMP 1319 (115)
C. phaseolus SAG 2013 (115)
Chroomonas sp. SAG 980-1 (115)
Teleaulax sp. SCCAP K-416 (115)
Cyanidioschyzon merolae 10D (108)
Cyanidium caldarium RK1 (119)
Cyanophora paradoxa (110)
Emiliania huxleyi (116)
Gracilaria tenuistipitata var. liui (121)
Heterosigma akashiwo CCMP 0452 (99)
Odontella sinensis (115)
Phaeodactylum tricornutum (115)
Porphyra purpurea (117)
Porphyra yezoensis (117)
Thalassiosira pseudonana (115)
Synechocystis sp. PCC 6803 (117)
The second conserved region (outside)
ATP/GTP-binding site (inside)

Fig.3.8.2 *(continued): The annotating for the conserved regions and hydrophilic and hydrophobic regions of* chl*I protein were based on codon-alignment of sequences from cryptophytes, Cyanobacteria and some red-line plastids.*

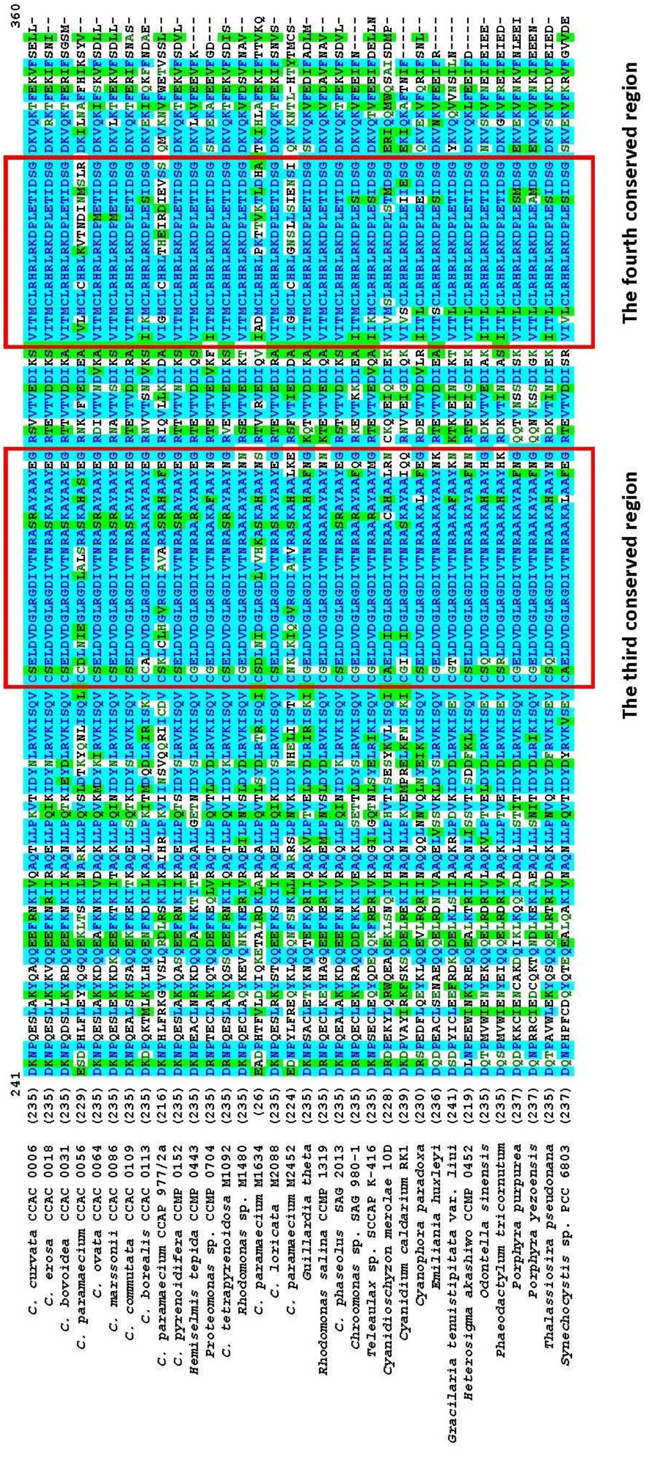

241
360
C. curvata CCAC 0006 (235)
C. erosa CCAC 0018 (235)
C. bovoidea CCAC 0031 (235)
C. paramaecium CCAC 0056 (229)
C. ovata CCAC 0064 (235)
C. marssonii CCAC 0086 (235)
C. commutata CCAC 0109 (235)
C. borealis CCAC 0113 (235)
C. paramaecium CCAP 977/2a (216)
C. pyrenoidifera CCMP 0152 (235)
Hemiselmis tepida CCMP 0443 (235)
Proteomonas sp. CCMP 0704 (235)
C. tetrapyrenoidosa M1092 (235)
Rhodomonas sp. M1480 (235)
C. paramaecium M1634 (26)
C. loricata M2088 (235)
C. paramaecium M2452 (224)
Guillardia theta (235)
Rhodomonas salina CCMP 1319 (235)
C. phaseolus SAG 2013 (235)
Chroomonas sp. SAG 980-1 (235)
Teleaulax sp. SCCAP K-416 (235)
Cyanidioschyzon merolae 10D (228)
Cyanidium caldarium RK1 (239)
Cyanophora paradoxa (230)
Emiliania huxleyi (236)
Gracilaria tenuistipitata var. liui (241)
Heterosigma akashiwo CCMP 0452 (219)
Odontella sinensis (235)
Phaeodactylum tricornutum (235)
Porphyra purpurea (237)
Porphyra yezoensis (237)
Thalassiosira pseudonana (235)
Synechocystis sp. PCC 6803 (237)
The third conserved region
The fourth conserved region

Fig. 3.8.2 *(continued): The annotating for the conserved regions and hydrophilic and hydrophobic regions of* chl*I protein were based on codon-alignment of sequences from cryptophytes, Cyanobacteria and some red-line plastids.*

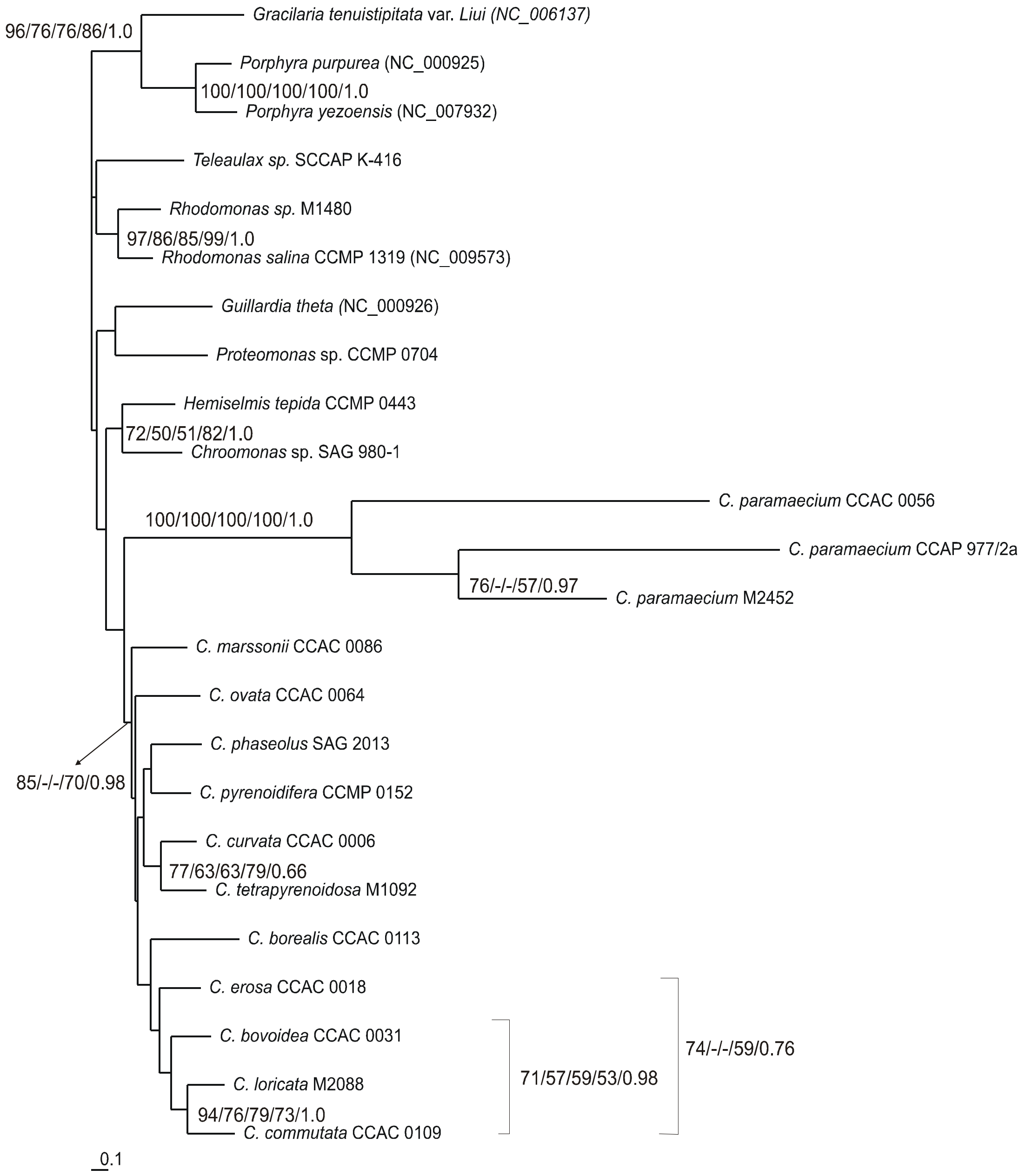


Fig. 3.8.3: *Rooted maximum-likelihood tree inferred from* chl*I gene sequences (24 taxa of 963 positions). The evolutionary model (GTR+I+G, - ln L= 14241.4952) were proposed by jModelTest 0.1 based on the corrected Akaike Information Criterion (AICc). From left to right: maximum likelihood bootstrap/maximum parsimony bootstrap/neighbor-joining bootstrap/logdet transformation bootstrap/posterior probabilities. Names without accession number: newly sequenced* chl*I in this study; scale bar = substitution per site.*

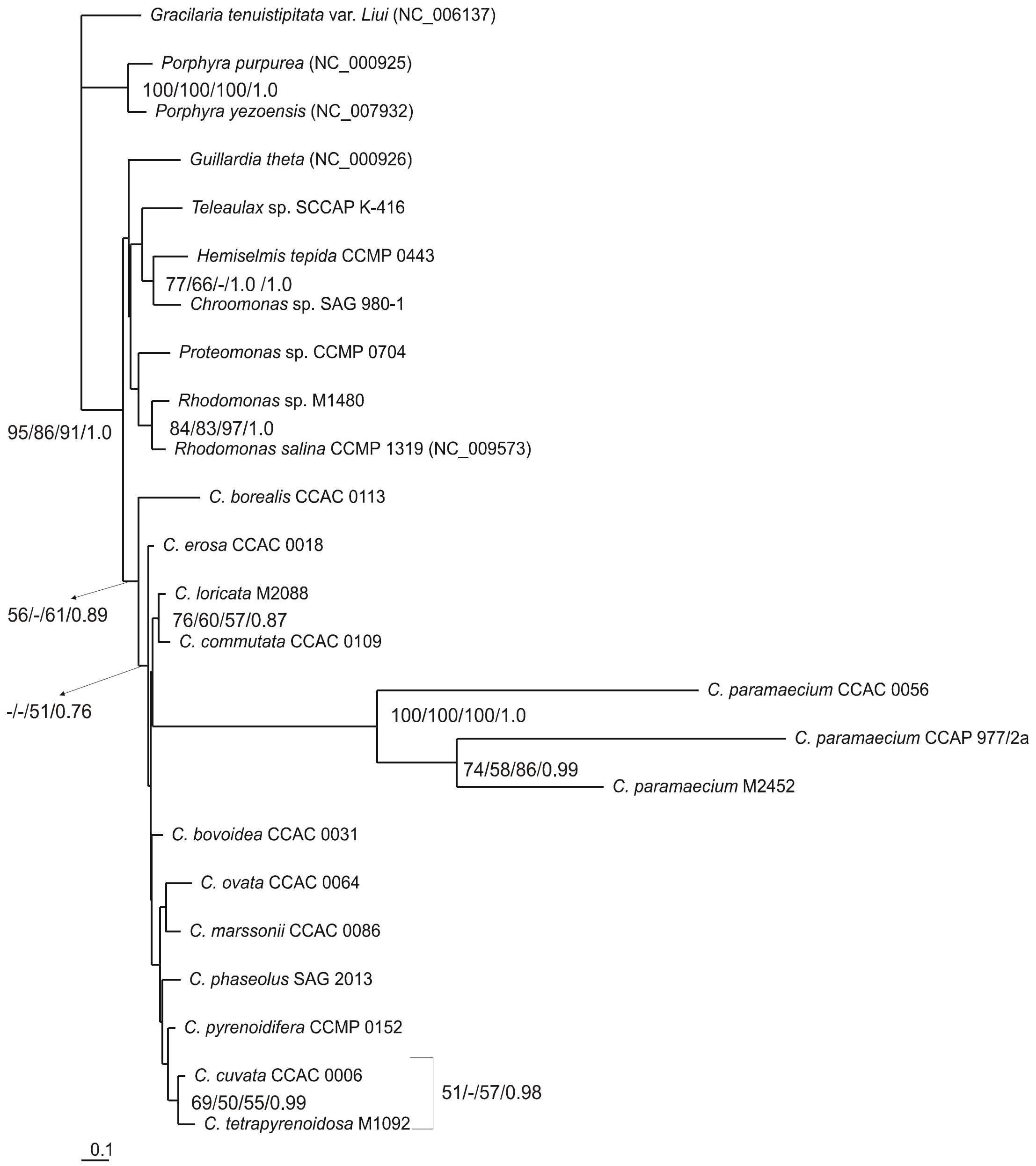


Fig. 3.8.4: *Rooted maximum-likelihood tree reconstructed based on 24* chl*I protein sequences (321 characters). The evolutionary model (CpRev+G, - ln L=5069.26) were selected according to the results of the corrected Akaike Information Criterion (AICc) in ProtTest 1.4 without "+F" option. Bootstrap values were assigned from left to right: maximum likelihood/maximum parsimony/neighbor-joining/posterior probabilities. Names without accession number: newly sequenced* chl*I in this study; scale bar = substitution per site.*

3.9 Analyses based one the cryptophyte plastid *rps*4 gene

3.9.1 Annotate conserved domains

To evaluate the conservation of the functional domains of *rps*4 protein, the data set mentioned above was extended to 36 taxa with the additional 12 taxa same as those in *chl*I database.

The result revealed the strictly conservation of secondary structure of *rps*4 protein regardless of photosynthetic or non-photosynthetic strains (Fig. 3.9.1). Moreover, the number and order of α-helices and β-sheets in red-line algae were not altered in comparison with *rps*4 of green algae or land-plant (Davies *et al.*, 1998 and Markus *et al.*, 1998). The amino acid residues involving in the important functions of *rps*4 such as putative RNA-binding, hydrophobic core, *etc*. seemed to be significantly constraint. However, the surprise came from *rps*4 protein of *Cyanidioschyzon merolae* and *H. akashiwo* CCMP 0152: the *rps*4 from the primary plastid lost completely the α-6 and a part of 3/10 helix while the latter did not contain the loop 10 between α-6 and β-4.

3.9.2 Reconstruct the phylogenetic trees

Like the *chl*I data set for phylogenies, three rhodophytes (*P. pupurea*, *P. yezoensis* and *G. tenuistipitata* var. *liui*) were kept as out-group. Codon-gaps at any positions were excluded prior any analyzes. The modified data set then had 597 nucleotide or/and 199 amino acid positions.

Table 3.9.1 showed the results of base composition calculating for entire taxa as well as the clusters of cryptophytes. In the entire data set, the third codon position got a high values of AT content (78.1%) while the first and second position contained a moderate values (52.2% and 60.7% respectively) suggested that the codons ending with AT would be dominant in the *rps*4 amino acid population. Like *chl*I gene observed above, *rps*4 gene had extremely high variable site number at the third codon positions, then followed by the first and second positions (196/199, 131/199 and 83/199, respectively). The large portion of parsimony informative sites in all codon positions (84%, 61.4% and 99% of variable sites for first, second and third codon positions, respectively) indicated that all position contributed to sense substitutions. These values were nearly the same when three out-group strains were removed from the data set.

Despite of the high abundant of AT content at the third codon position, the result of test for stationarity of base composition with the Chi-square test implemented in PAUP* v4.0 showed that the data set passed the test (Chi-square = 47.698593, df=78, P = 0.99731159).

To select the best-fitting evolutionary model for *rps*4 gene and *rps*4 protein database proposed, jModeltest 0.1.1 and ProtTest 1.4 were employed. The TIM3+G and cpRev+G model were proposed as the selected models, respectively.

Model selected:

Model = TIM3+G
partition = 012032
-lnL = 9696.0158
K = 59
freqA = 0.4048
freqC = 0.1218
freqG = 0.1331
freqT = 0.3403
R(a) [AC] = 11.0136
R(b) [AG] = 23.6176
R(c) [AT] = 1.0000
R(d) [CG] = 11.0136
R(e) [CT] = 32.7261
R(f) [GT] = 1.0000
gamma shape = 0.2700

In the rooted maximum likelihood tree of *rps*4 nucleotide sequences, the *Cryptomonas* strains separated into two main clades without support (Fig. 3.9.2).

The first main clade contains the colorless M1634 and CCAC0018, CCAC0031, CAC0109 and M2088 in which the colorless strain occupied at the early divergence position with slight increased evolutionary rate. The cluster CCAC0109-M2088 had the support values from moderate to high (71%/64%/53%/0.94) while the cluster

CCAC0108-CCAC0031 got the bootstrap values from maximum likelihood and posterior probability analyzes only (62%/-/-/0.98).

The remaining *Cryptomonas* strains were pushed into the second main clade in which they were further subdivided into two small clades. The three colorless strains grouped with CCAC0107, CCAC0108, CCAC0113 and SAG 2013 to form a clade that was nearly the same with long-branch clade observed in the *rbc*L gene tree. However, the cluster CCAC0107-CCAC0108 became the closest relative to three colorless in the *rps*4 gene tree instead of CCAC0113 in the *rbc*L gene tree.

Unlike the *chl*I protein phylogenetic tree, the maximum likelihood tree of *rps*4 protein sequences was nearly congruent with *rps*4 nucleotide sequence tree (Fig. 3.9.3). Generally, all *Cryptomonas* strains were distributed into two main halves, each had the same terminal nodes with those of *rps*4 gene tree. However, the positions of some strains were changed. For example, the colorless M1634 offered the basal position to strain M2088; while strain SAG 2013 split out of CCAC0013 to become close sister with three colorless (but without support).

Observation the evolutionary rate of colorless lineages showed that they had a slightly accelerated evolutionary rates in the *rps*4 gene trees while these rate were significantly increased in the protein trees.

3.9.3 Codon usage

Table 3.9.2 displayed the dominant numbers of NNU in compassion to NNC in two-fold degenerate NNY codon of cryptophytes *rps*4 gene.

Table 3.9.1: Base composition and variability of *rps*4 sequences

	ALL TAXA (27 TAXA)								CRYPTOPHYTES (24 TAXA)			
	T%	C%	A%	G%	AT%	Conserved sites	Variable sites	Parsimony-Informative sites	Conserved sites	Variable sites	Parsimony-Informative sites	Total
1st codon position	19.4	20.1	32.8	27.7	52.2	68	131	110 (84% of variable sites)	79	120	103 (88.8%)	199
2nd codon position	28.9	18.3	31.8	21.0	60.7	116	83	51 (61.4%)	126	73	42 (58%)	199
3rd codon position	35.9	8.6	42.2	13.2	78.1	3	196	194 (99%)	4	195	190 (97.4%)	199
All sites	28.1	15.7	35.6	20.6	63.7	187	410	355 (86.6%)	209	388	335	597

Table 3.9.2: Codon usage of NNC and NNY codon in cryptophytes *rps*4 gene

	STRAINS	UAC/UAU TYR	CAC/CAU HIS	AAC/AAU ASN	GAC/GAU ASP	UGC/UGU CYS	UUC/UUU PHE
01	CCAC 0006 (M0420)	3/5	0/4	4/7	4/4	1/1	1/2
02	CCAC 0086 (M1473)	**5/3**	0/4	3/8	0/4	1/0	1/3
03	CCAC 0064 (M0847)	2/7	0/4	3/6	1/4	0/2	0/3
04	CCMP 0152	4/4	0/4	2/9	1/3	0/2	0/3
05	M1092	2/6	0/4	2/10	1/3	0/2	**2/1**
06	CCAC 0031 (M1094)	1/7	0/4	1/9	2/3	0/2	**2/1**
07	CCAC 0109 (M0739)	3/5	2/2	1/9	2/3	0/2	0/2
08	CCAC 0018 (M0788)	4/4	1/3	1/8	0/5	1/1	0/3
09	M2088	2/6	0/4	2/7	1/3	0/2	0/3
10	SAG 2013	3/5	0/4	4/6	1/4	0/2	1/2
11	CCAC 0108 (M1079)	2/6	0/4	4/6	1/5	0/2	0/3
12	CCAC 0107 (M0850)	0/7	0/4	2/8	2/4	0/1	0/3
13	CCAC 0113 (M1083)	1/7	2/3	3/8	0/3	0/1	0/3
14	CCAP 977/2a	2/6	1/6	1/6	1/5	1/0	1/5
15	CCAC 0056 (M1303)	2/5	1/5	2/7	0/4	0/1	0/5
16	M2452	0/7	2/3	2/8	0/2	2/1	0/7
17	M1634	1/4	**4/3**	3/4	0/4	0/2	1/5
18	SAG 980-1	2/6	2/2	0/11	0/5	0/1	1/3
19	CCMP 0443	2/6	3/1	2/9	2/3	0/2	2/2
20	CCPM 0704	3/5	1/3	4/5	1/1	**1/0**	1/2
21	M1480	3/5	1/3	4/6	2/5	0/2	1/2
22	SCCAP K-416	2/6	2/2	1/6	1/5	0/2	2/3
23	*G. theta*	**5/3**	1/4	3/8	1/4	1/0	1/2
24	CCMP 1319	**5/3**	1/3	3/8	0/6	0/2	**2/1**

The rps4 data set used in phylogenetic analysis was used for counting NNY codon per 199 codon (597 nucleotide positions); bold face values indicate some exception in which NNC over NNT codon.

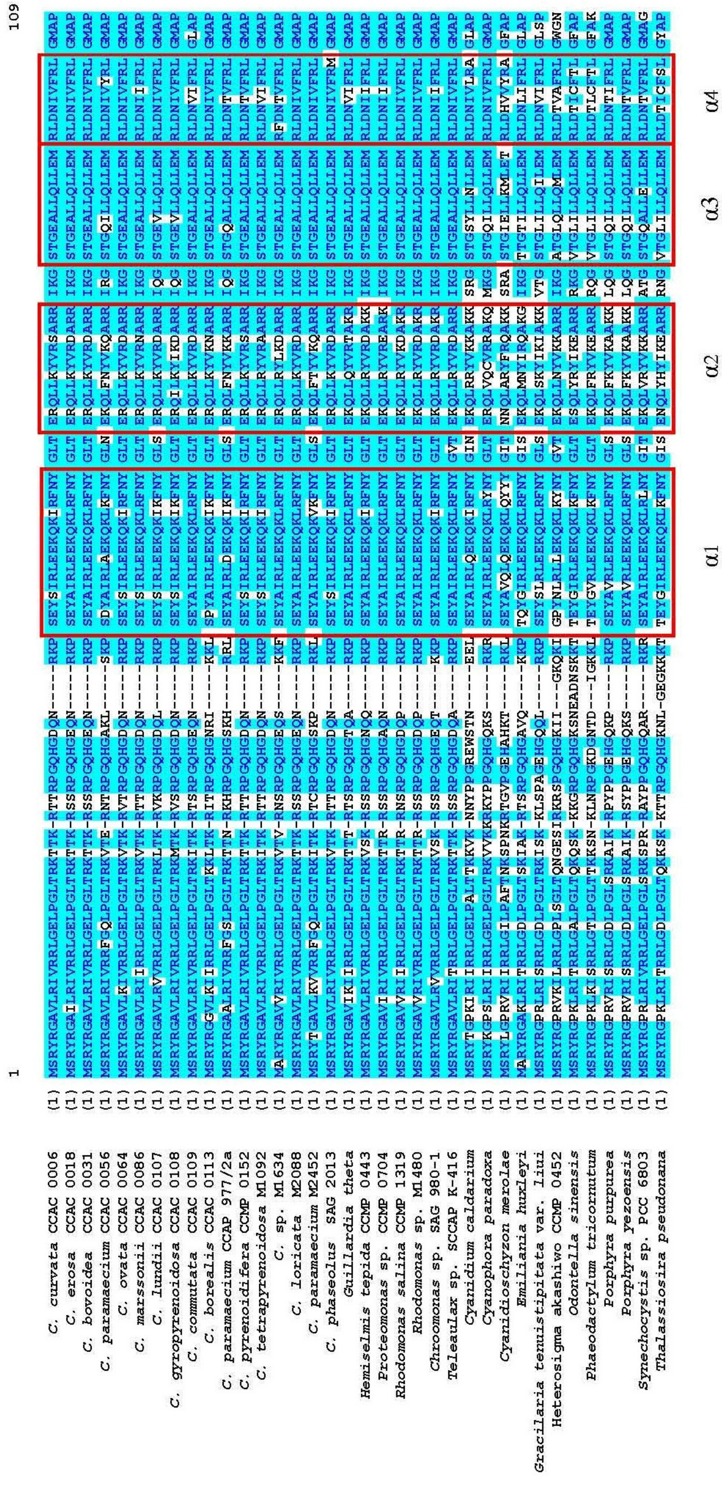


Fig. 3.9.1: *Codon-alignment of* rps*4 protein sequences from cryptophytes, Cyanobacterium and some red-line plastids showed the numbers and order of α-helices and β-sheets.*

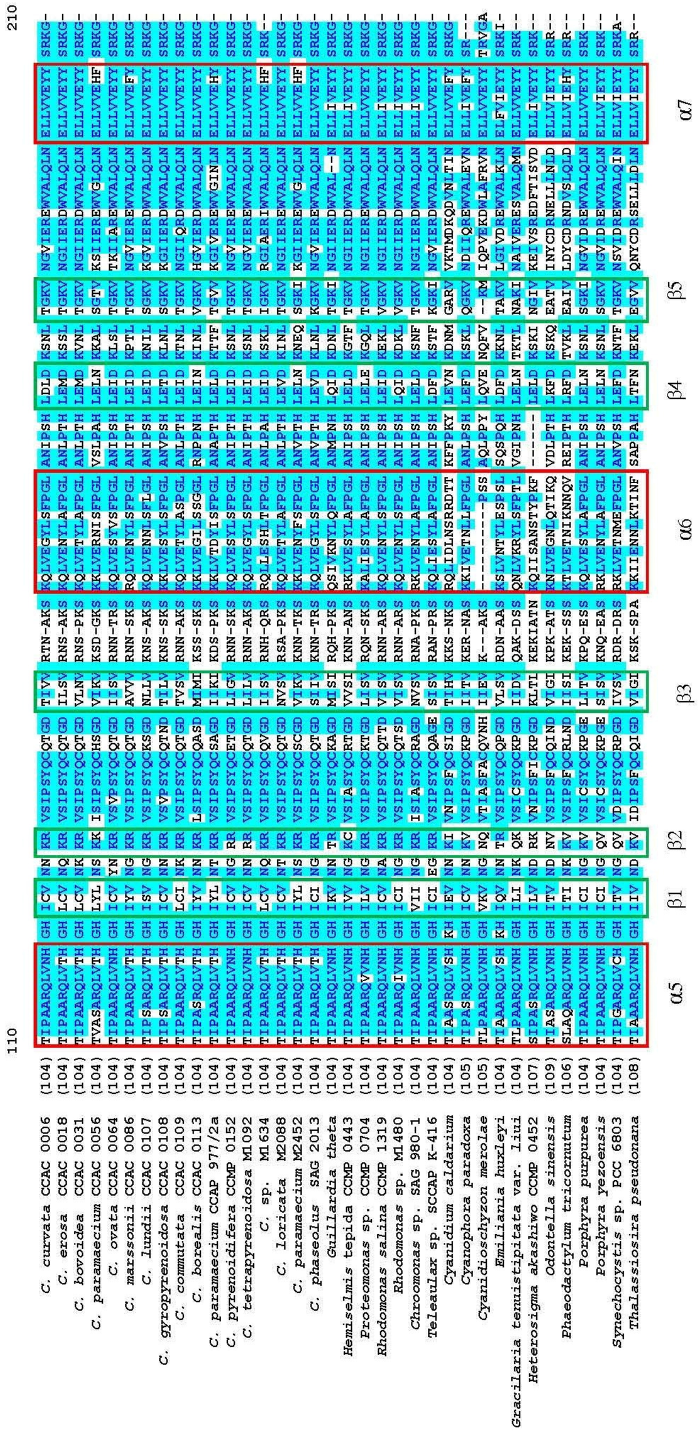


Fig. 3.9.1 *(continued): Codon-alignment of* rps*4 protein sequences from cryptophytes, Cyanobacterium and some re-line plastids showed the numbers and order of α-helices and β-sheets.*

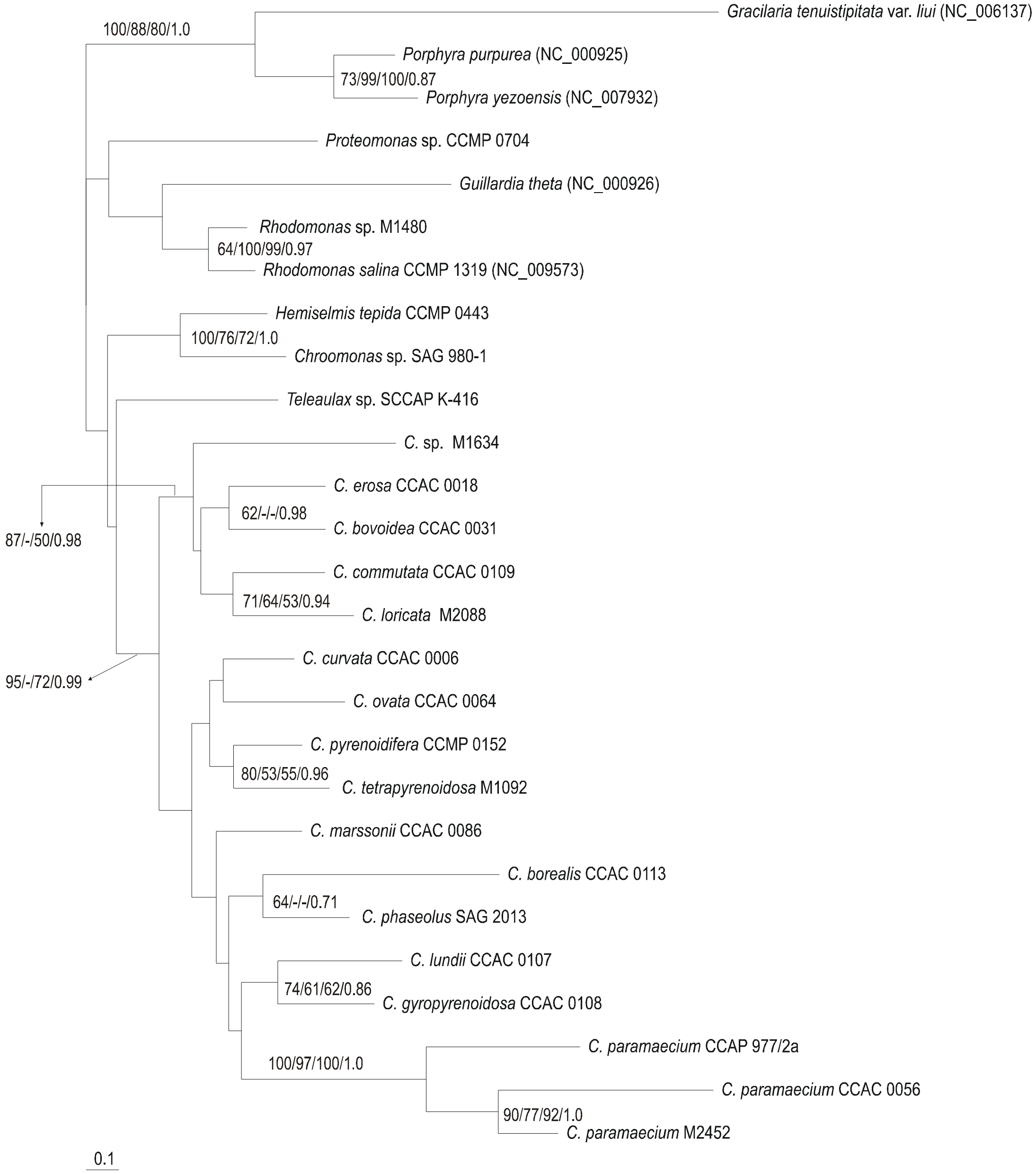


Fig. 3.9.2: *Rooted maximum-likelihood tree inferred from* rps*4 gene sequences (27 taxa of 597 positions). The evolutionary model (TIM3+G, - ln L= 9696.0158) were proposed by jModelTest 0.1 based on the corrected Akaike Information Criterion (AICc). From left to right: maximum likelihood bootstrap/maximum parsimony bootstrap/neighbor-joining bootstrap/posterior probabilities. Names without accession number: newly sequenced* rps*4 in this study; scale bar = substitution per site.*

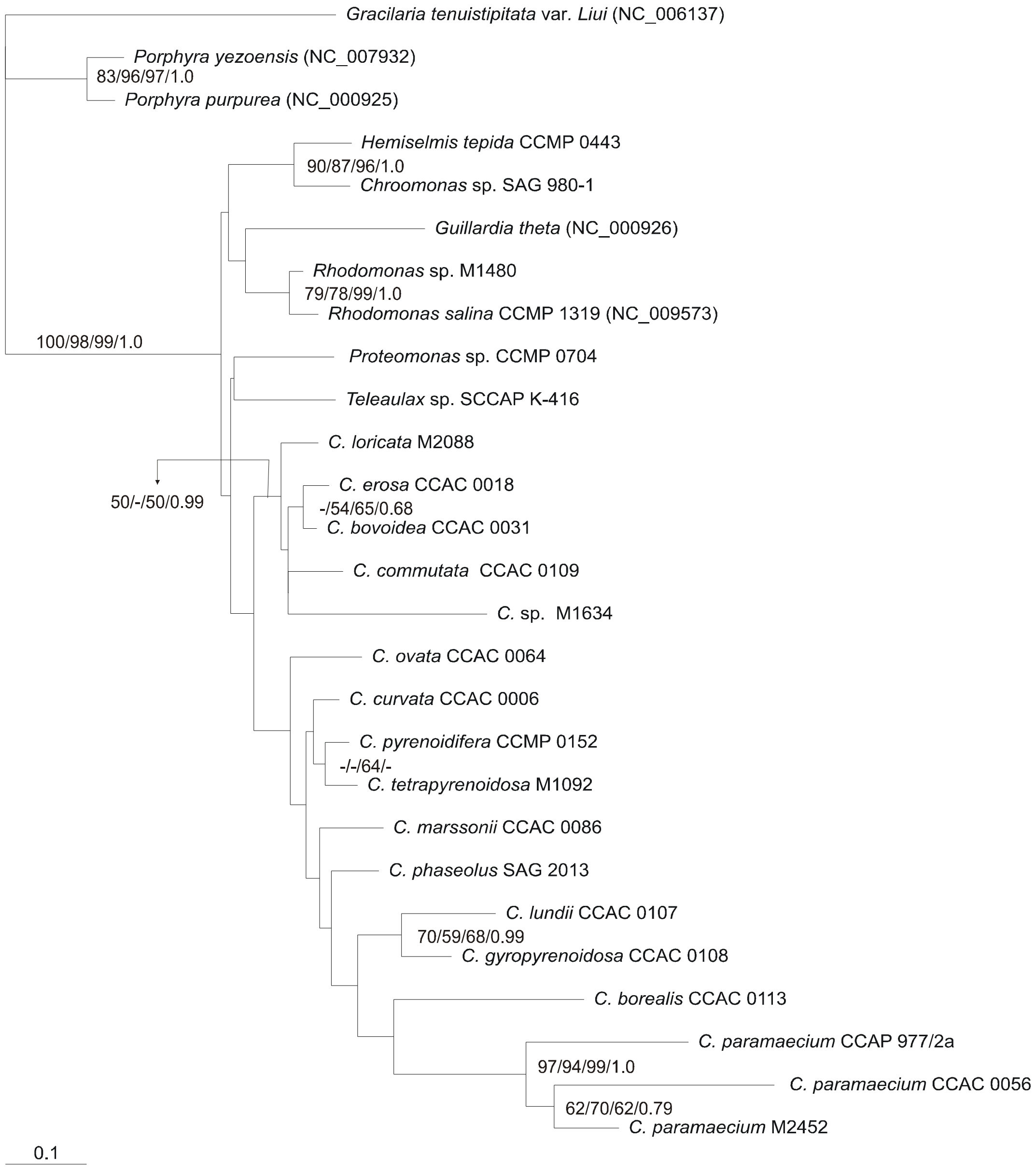


Fig. 3.9.3: *Rooted maximum-likelihood tree reconstructed based on 27* rps4 *protein sequences (199 characters). The evolutionary model (CpRev+I+G, - ln L=3250.31) were selected according to the results of the corrected Akaike Information Criterion (AICc) in ProtTest 1.4 without "+F" option. Bootstrap values were assigned from left to right: maximum likelihood/maximum parsimony/neighbor-joining/posterior probabilities. Names without accession number: newly sequenced* rps4 *in this study; scale bar = substitution per site.*

3.10 Analyses based on the concatenated data sets of *chl*I, *rps*4 and *rbc*L genes

3.10.1 Reconstruct phylogenetic trees

Base on the separated data sets that have done phylogenies in previously sections, a concatenated data set was established include *chl*I gene (963 nucleotide positions), *rps*4 gene (597 nucleotide positions) and *rbc*L gene (972 nucleotide positions). *P. pupurea*, *P. yezoensis* and *G. tenuistipitata* var. *liui* still involved as out-group while strains M1634, CCAC0107 are CCAC0108 did not participate in as they lack one or two genes above.

The entire concatenated data set was not passed the chi-squares test for homogeneity of base frequencies across taxa (Chi-square = 115.738636, df=69, P = 0.00036474). However, the P-value increased to 0.99938197 when the third co don positions were removed (Chi-square = 37.206271 df=69, P = 0.99938197). This was congruent with *chl*I and *rbc*L data sets as they only passed the chi-squares test after exclusion of the third codon positions.

As did for smaller database, the combined data set was subjected to jModeltest 0.1.1 and ProtTest 1.4 to detect the best-fitting evolutionary model. Two models TIM3+I+G and CpREV+I+G were selected for nucleotide and protein sequences, respectively.

```
Model selected:
 Model = TIM3+I+G
 partition = 012032
 -lnL = 32546.6433
 K = 54
 freqA = 0.3828
 freqC = 0.1222
 freqG = 0.1397
 freqT = 0.3553
 R(a) [AC] = 4.2253
 R(b) [AG] = 8.2980
 R(c) [AT] = 1.0000
 R(d) [CG] = 4.2253
 R(e) [CT] = 18.1463
 R(f) [GT] = 1.0000
```

p-inv = 0.1960
gamma shape = 0.3500

The rooted maximum likelihood tree showed that the *Cryptomonas* strains divided into two main halves (Fig. 3.10.1 and Fig. 3.10.2). The first clade contained CCAC 0018, CCAC 0031, CCAC 0109 and M2088 with high bootstrap values (99%/89%/92%/99%/1.0) while the second clade comprised 10 remaining strains without bootstrap support in which CCAC 0013 and three heterotrophic strains formed a separated clade with moderate bootstrap values (52%/78%/85%/75%/0.77). Due to the lack of strains CCAC 0107 and CCAC0108 in this concatenated data set, the presence of long-branch was still questionable. However, the evolutionary rates of 3 colorless and CCAC 0113 were higher than other *Cryptomonas.* Three strains CCAC0006, M1092 and CCMP0152 again displayed the attachment to each other even though the bootstrap value was not so high (61%/72%/57%/92%/0.73).

In the amino acid translated data set, the three colorless *Cryptomonas* still displayed an accelerated evolutionary rate as well as early divergent position in *Cryptomonas* clade (Fig. 3.10.3). Strain CCAC 0113 also had a high evolutionary rate but not grouped with three colorless as seen in nucleotide tree, it otherwise located at the basal position of one of two main clades without support. The other strain, includes non-*Cryptomonas* strains, branched similar to those of nucleotide phylogenetic trees.

3.10.2 Calculate distance genetics

Maximum likelihood distances between strain CCMP0152 and other cryptophytes were calculated in each genes *chl*I, *rps*4 and *rbc*L. They then were transferred to chart diagram (Fig. 3.10.3). Obviously, the maximum likelihood distance increased significantly in three colorless strains in all data sets, especially, extremely high substitution rate in *chl*I data set was observed. Apparently, the evolutionary rates increased from *rbc*L to *rps*4 and *chl*I genes in cryptophyte strains.

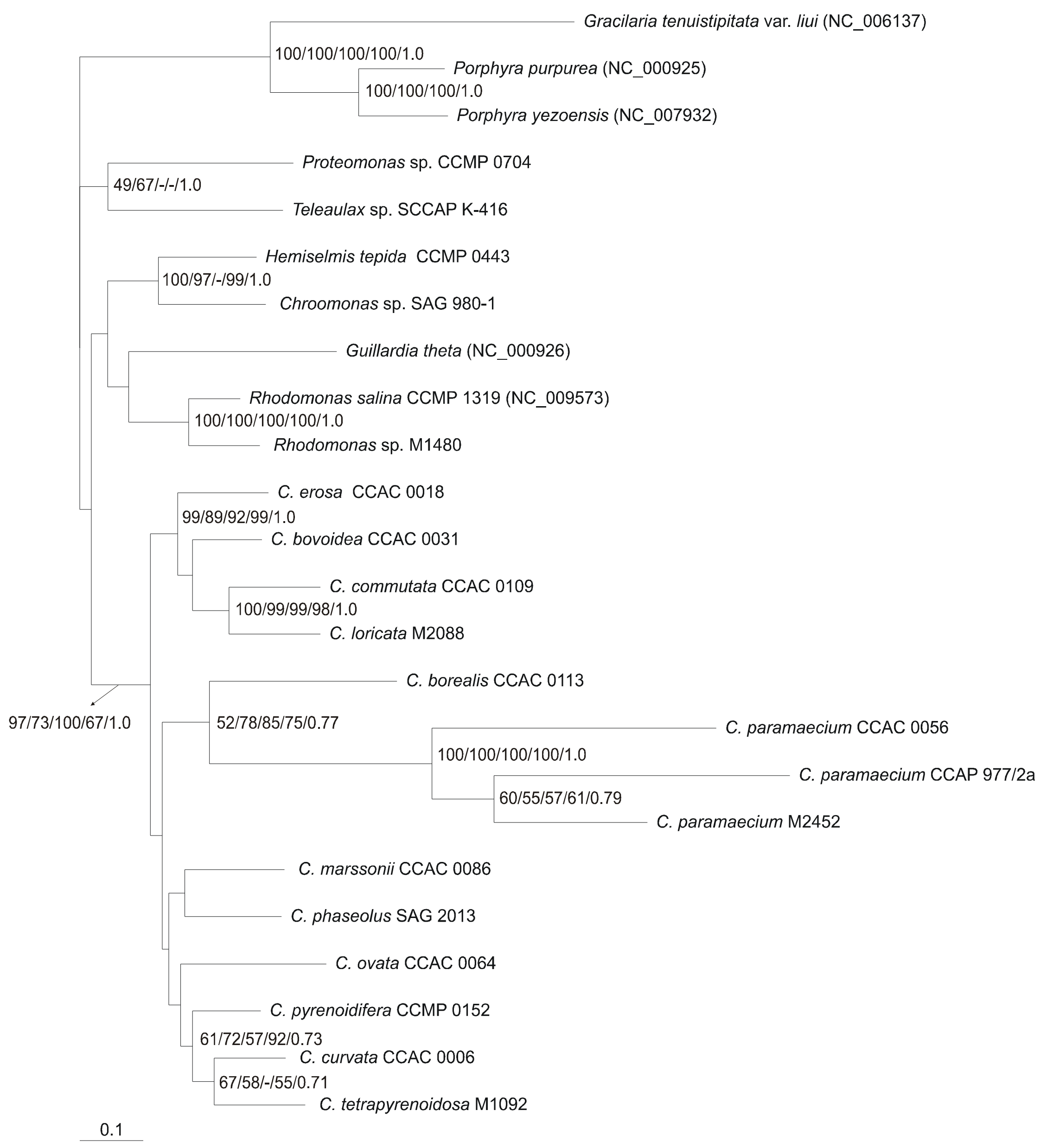


Fig. 3.10.1: *Rooted maximum-likelihood tree inferred from concatenated plastid* chl*I,* rps*4 and* rbc*L gene sequences (24 taxa of 2532 positions). The evolutionary model (TIM3+I+G, - ln L= 32546.6433) were proposed by jModelTest 0.1 based on the corrected Akaike Information Criterion (AICc). From left to right: maximum likelihood bootstrap/maximum parsimony bootstrap/neighbor-joining bootstrap/logdet transformation bootstrap/posterior probabilities; scale bar = substitution per site.*

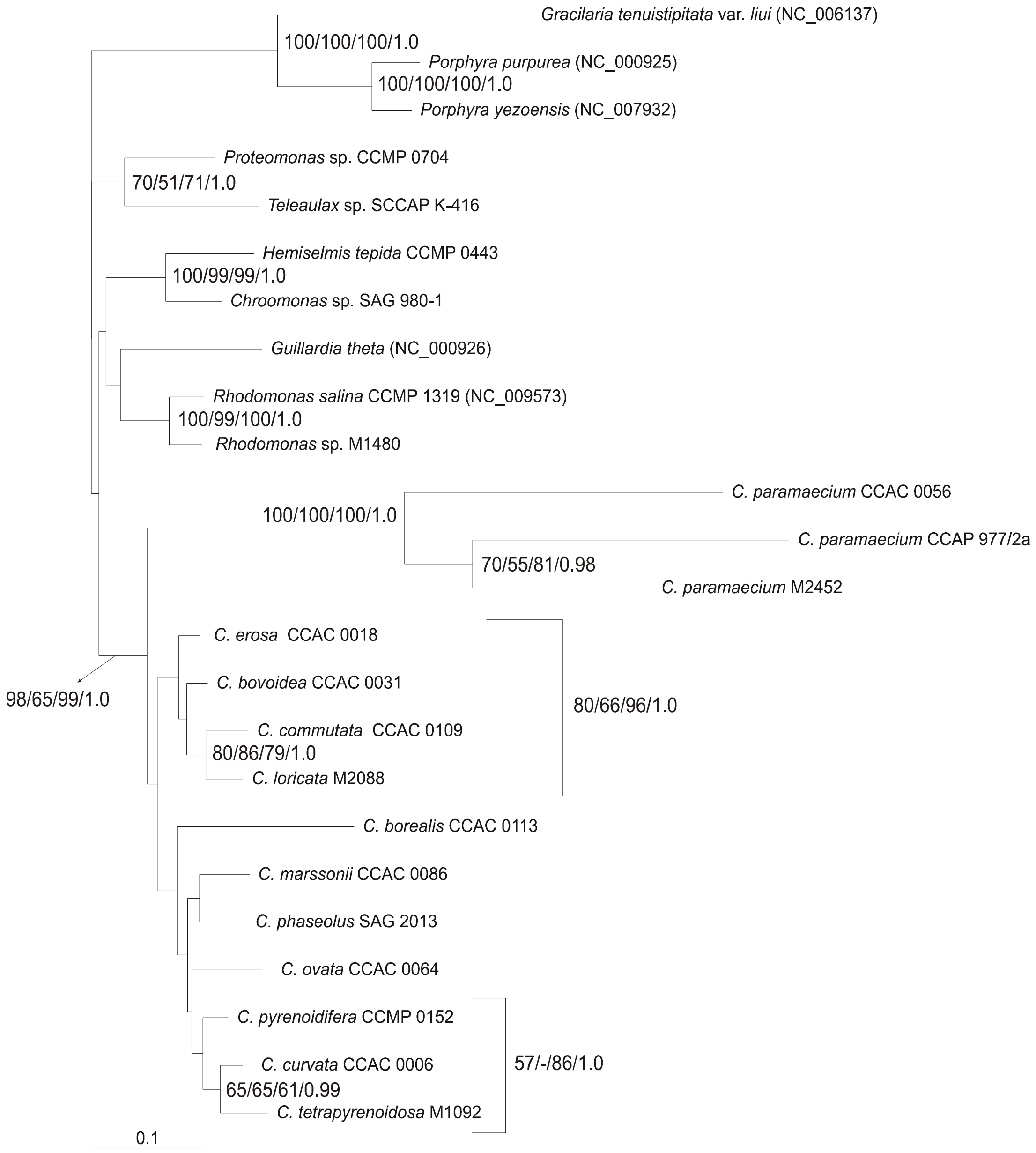

Gracilaria tenuistipitata var. *liui* (NC_006137)
100/100/100/1.0
Porphyra purpurea (NC_000925)
100/100/100/1.0
Porphyra yezoensis (NC_007932)
Proteomonas sp. CCMP 0704
70/51/71/1.0
Teleaulax sp. SCCAP K-416
Hemiselmis tepida CCMP 0443
100/99/99/1.0
Chroomonas sp. SAG 980-1
Guillardia theta (NC_000926)
Rhodomonas salina CCMP 1319 (NC_009573)
100/99/100/1.0
Rhodomonas sp. M1480
C. *paramaecium* CCAC 0056
100/100/100/1.0
C. *paramaecium* CCAP 977/2a
70/55/81/0.98
C. *paramaecium* M2452
C. *erosa* CCAC 0018
C. *bovoidea* CCAC 0031
98/65/99/1.0
80/66/96/1.0
C. *commutata* CCAC 0109
80/86/79/1.0
C. *loricata* M2088
C. *borealis* CCAC 0113
C. *marssonii* CCAC 0086
C. *phaseolus* SAG 2013
C. *ovata* CCAC 0064
C. *pyrenoidifera* CCMP 0152
C. *curvata* CCAC 0006
57/-/86/1.0
65/65/61/0.99
C. *tetrapyrenoidosa* M1092
0.1

Fig. 3.10.2: *Rooted maximum-likelihood tree reconstructed based on a concatenation of* chl*I,* rps*4 and* rbc*L protein sequences (24 taxa of 844 characters). The evolutionary model (CpRev+I+G, - ln L=10619.92) were selected according to the results of the corrected Akaike Information Criterion (AICc) in ProtTest 1.4 without "+F" option. Bootstrap values were assigned from left to right: maximum likelihood/maximum parsimony/neighbor-joining/posterior probabilities; scale bar = substitution per site.*

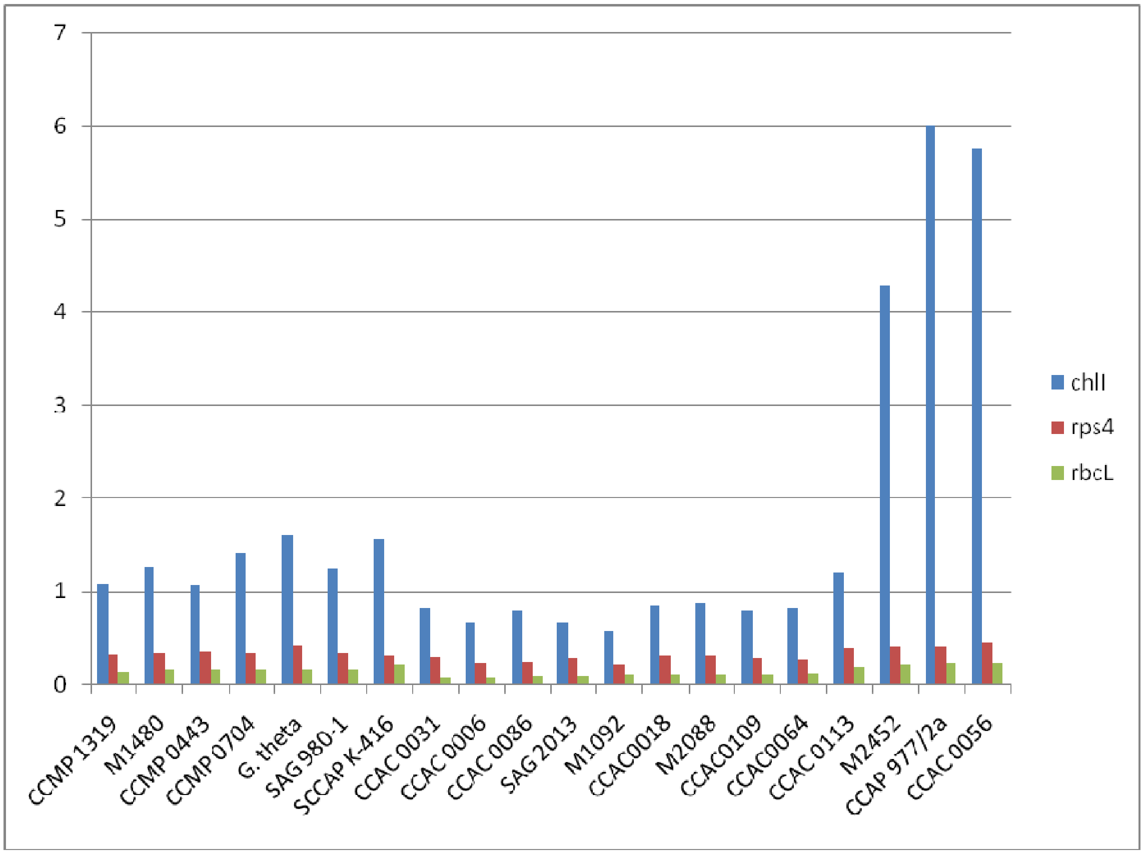


Fig. 3.10.3: *The maximum likelihood distances (genetic distances) among cryptophytes with three data sets, strain* C. pyrenoidifera *CCMP 0152 was used as reference.*

4. DISCUSSIONS

4.1 Advantages and disadvantages of long-range PCR and primer-walking sequencing combination

Though there were some gaps in several strains need to be full-filled, the long-range PCR combined with primer-walking methods applied to purify and read the nucleotide compositions of the 16S rRNA–*rbc*L fragments of wide range cryptophyte strains simultaneously was successful.

The 23S rRNA–*rbc*L fragment was chosen for reading as it comprised ribosomal operon, transfer RNA and protein coding genes, which have very high conservative degrees, satisfying the strictly criterions for constructing the primers for long-range PCR. Alternatively, the designing for primers for long-range PCR was facilitated by the published plastid ribosomal operon and *rbc*L sequence database as well as newly readings. Therefore, all primers for long-range PCR were universal and were expected to be applied successfully for other cryptophytes.

The long-range PCR was almost successful for all examined strains thanks to the carefully prepared optimization of PCR protocol, except for the colorless M1634 strain that only the fragment from 16S rRNA to *rps*4 gene was amplified.

The most advantage of the used approach was that not so many initial DNA materials were required; about 10 to 25 ng of DNA template was satisfied for one reaction. This helped to save time and cost in preparation of DNA material. Unlike the isolation chloroplast by ultracentrifugation, dozens of grams of fresh biomass were needed to obtain the total DNA and subsequently to segment the genomes into separately bands in two or three days. In fact, the colorless strain *C.* SAG 977-2f was one of the favorite choices for searching and reading its *rbc*L as well as 16S rRNA–*rbc*L fragments. Unfortunately, the strain always died after few weeks arrived from SAG collection; and the available frozen total DNA was not enough for running analysis. The success of amplification a plastome large fragment by MasterAmp^{TM} Extra-long PCR kit will open an opportunity to be back for continuing the uncompleted lab-workings on this interesting strain.

The utilizing set of 45 genes retained in all plastid genomes for designing universal primers then combining them in many optimal manners to produce large fragments is like using endonuclease restriction enzymes techniques to cut genome into small pieces, subsequently mapping and reading whole plastome. However, less initial material was required as mentioned above and no more time is spent to read and infer the plastid map based on cut segments.

The shotgun approach seems to be the best choice nowadays for large-scale sequencing projects; but more time, cost, labor and high-tech instruments are required to apply this trend. The primer-walking strategy, on the contrary, requires standard molecular biology facilities only (Sterky & Lundeberg, 2000).

The ultracentrifugation methods, as Douglas *et al.* (2001) conceded, also caused the contamination of mitochondria genome into other bands by many factors. Amplifying the larger fragments of plastid genome by long-rang PCR avoided this unendurable issue.

The lab-workings of this project were done in half a year could be considered beneficial in comparison with other methods.

Inspire of these advantages, there are some unprejudiced problems:

- The *ycf*26 gene and ORF403 in all strains were highly divergent in both nucleotide compositions, sizes caused the sequencing primers for ORF403, and *ycf*26 genes were nearly useless. Several specie-specific sequencing primers for these two genes had to be done to full-fill the gaps.
- The absence of *psa*M as well as the substantial sequence divergence at the upstream portion of *chl*I gene in colorless lineages caused the same problems.

These disadvantages seemed not to be major problems, and were overcome easily.

In summary, the combination of long-range PCR with primer-walking approaches is reliable in the future not only to get the sequence database of the same fragments from another cryptophyte strains but also to be able to extent for whole plastome of examined *Cryptomonas*. The approach actually opened an ability to harvest huge amount of inter-species plastid genome information quickly and economically.

4.2 *psa*M and *chl*I formed a cluster in red-line plastids.

An attractive finding from NCBI Genome Map search pointed that *psa*M and *chl*I genes formed a tightly cluster in almost red-line plastomes. The clustering of genes in the plastome is not new phenomena. In the early stage of genome era, when the biological data from genomes of chloroplast or/and nucleus were still limited, the researchers had based on the clusters of genes recognized in relative examined genomes as one of the most important feature to deduce the origin and evolutionary history of these genomes (Ohta *et al.*, 1997). Observing from cyanobacterial genomes to plastid geomes, the researchers also discovered that some genes widely distributed in Cyanobacteria genomes but they would fused to form clusters in some plastid genomes from several lineages after the endosymbiosic process had taken place, for example ribosomal protein cluster, the rRNA cistrons, the *rpo*BC/*atp*A cluster, and the *psb*BTNH cluster (Douglas & Penny, 1999). Another example that researcher successful exploited the phylogenetic information from gene clusters in chloroplast genomes was that *G. theta* plastome directly originated from red-lineage plastid thought out the comparison of the distribution, arrangements of gene clusters in full-sequenced plastome *G. theta*, *P. pupurea*, *C. paradoxa*. This hypothesis afterward had been supplemented, developed and accepted widely when the abundant biological information easy obtained from plastomes as well as the assistance of more powerful analyze tools (Reith & Munholland, 1993).

As mentioned above, the *psa*M functions have not been established as well as the signal of its protein and mRNA have not yet been proved (Nelson & Yocum, 2006). Scanning the 5' non-coding region of *psa*M or/and *chl*I showed that most regulation sites such as SD sequences, boxes -10 and -35 were found in front of *psa*M gene while nothing was detected in the upstream part of *chl*I gene except for an adenine-rich block (Fig. 3.5.2 and Fig. 3.5.3). This suggested that the pasM-*chl*I cluster seemed to be a co-transcriptional unit (di-cistronic unit) but the translational regulation for *chl*I gene could be directed by another pathway in which adenine-rich region to be involved.

4.3 The first evidence of reduced size of cryptophyte leukoplast genomes

Surprisingly, no *psa*M gene was found in the colorless lineage of CCAP 977/2a, M2452 and CCAC 0056. There are two possibilities to judge the fate of *psa*M in leukoplast-bearing cryptophytes: the gene must have been lost completely or relocated to somewhere in their plastomes. It was assumed that the translocating of *psa*M, had it happened in colorless *Cryptomonas* strains, should be coupled with its partner *chl*I. In this case, *chl*I would not be found in between 16S rRNA and tRNA-R genes. In fact, *chl*I gene was found in the 16S rRNA–*rbc*L fragments of colorless plastomes. Therefore, the possibility of *psa*M translocation seems not to be persuasive.

Moreover, unlike the main genetic machinery located in nucleus, the chloroplast genomes are non-recombination and uniparental inheritance, hence the translocation of certain gene(s) in plastome(s) of interspecies are almost not occurring. The evolutionary routine of colorless plastids prefered to reject unnecessary genes and compressed their plastomes rather than rearranged their genes (Sato, 2006). These hypothesis supported the assumption that *psa*M was lost completely in heterotrophic cryptophytes instead of being transported to another position in leukoplast genome of *C. paramaecium*.

Even thought whole plastid genome sequencing of colorless cryptophyte strains have not yet been done completely to compare with other photosynthetic relatives, the comparison of 16S rRNA–*rbc*L fragment sizes showed that the leukoplast plastids of cryptophytes seemed to reduce their genome sizes which was evinced not only in the gene loss (*psa*M, *ycf*26 gene) but also in the curtailment of the non-coding regions and overlapped genes. The result is preliminary evidence that strongly support the hypothesis of plastome size reduction in colorless *Cryptomonas* lineages that was revealed by Hoef-Emden (2005a).

4.4 First overviews lock about ORF403

Still now, the biological role of Tic22 (ORF403) protein in protein import apparatus at the chloroplast membrane was affirmed evidently in some published papers. However, most of them only issued the recognized biological role of that protein; no more papers have presented the deeply investigation of relationship between structure and function of Tic22 or phylogenies of this interesting object.

The paper of Kalanon & McFadden (2008), in a comprehensive bioinformatic examination of the chloroplast protein translocation complex, pointed out some characteristics of Tic22 protein, however, only in green algae, land plants and some Cyanobacteria.

This study, not only published some newly sequenced ORF403 of cryptophyte plastomes but also built an extended Tic22 collection of 70 sequences distributed in many kinds of genomes, opened the first look about the evolution of Tic22 protein after it diverged from its Cyanobacterial ancestors. The codon- alignment of Ti22 amino acid sequences revealed several specific characteristics that only were found in cryptophyte and red-line plastids.

4.5 *ycf*26 in cryptophyte plastids seems to be a pseudogene than alters function

Martin *et al.* (1998), Ashby *et al.* (2002) and Martin *et al.* (2002) showed that the *ycf*26 gene was deleted from all green-line plastome in very early stage of evolution history of these organisms while this gene was still found in some red-lineage plastomes such as Rhodophyta, Haptophyta and Heterokontophyta. The finding of *ycf*26 several Cryptophyta species as well as in some *Cryptomonas* strains in this study showed that the non-universal distribution of this gene was not only at the phylum level but also in species and lineage level

Analyzing the functional domain of current *ycf*26 pointed out that the *ycf*26 proteins were divided into two groups in which they were different in the numbers of the domains recognized in the *ycf*26 proteins. The first group (mostly red-algae and one representative for haptophyte) held the numbers of functional domains like those of Cyanobacteria while the second group (Cryptophyta and Raphidophyta) mostly contained the HiKa and HATPase_c domains in which evidence for the degradation of the HATPase_c domains were emerged.

Whether the truncated-*ycf*26 proteins in Cryptophyta as well as Raphidophyta still have had the functions like those of full–*ycf*26 protein in Cyanobacteria or their biological functions were altered to compatible with their current structures or they were relic of full *ycf*26 on the going to become pseudogenes?

In the activity model of *ycf*26 deeply investigated in Cyanobaceria (Stock *et al.*, 2000; Suzuki *et al.*, 2000; Mikami *et al.*, 2002; van Waasbergen *et al.,* 2002; Mikami *et al.,* 2003; Morrison *et al.*, 2005; Ashby & Houmard, 2006; Morici *et al.*, 2006; Kanesaki *et al.*; 2007), the PSD domain was assigned as the domain responsible for receiving the extracellular stimuli, the cooperation between PSD and PAS domains was assumed to enable *ycf*26 to sense a wide variety of stress conditions while the HAPTase_c played a very important role in the support energy and catalyze the signal transmission from sensory histidine protein kinase to the response regulator protein.

The truncated-*ycf*26 in Cryptophyta and Heterokontophyta lost completely the stimulus receiving domains (PSD and PAS) as well as having degraded the catalyze HATPase_c domain. Therefore, it is hardly to answer how truncated-*ycf*26 can play its function with only the HiKA domain?

Lerat & Ochman (2005) demonstrated that one of the feature of pseudogenes that they have the moderate AT% values in comparison with functional genes and non-coding regions. The calculate the AT% for non-coding region in the cryptophyte 16S rRNA – *rbc*L fragments containing *ycf*26 showed that the AT% of *ycf*26 complied with this rule (Table 3.4.1). The extended calculation for *ycf*26 occupied in full sequenced plastid genomes also obtained the same results: the AT% contents of these *ycf*26 were always higher than AT% content of their plastomes. In contract, most AT% contents of *ycf*26 in cyanobacterial genomes were lower than A T% content of their host genomes (see Appendix 22). Obviously, the increasing AT% content over the AT %

content of its plastomes could be considered indicative for genome compositional bias as well as the *pseudogenezation*.

Duplessis *et al*., (2007) assumed that the *ycf*26 gene of *H. akashiwo* CCMP 0152 was split into several small parts. However, it was not supported by detailed analysis of the completely sequenced *H. akashiwo* CCMP 0452 chloroplast genome. In the same way, the traces of trans-membranes, sensor and HAMP domains were not detected in the full sequenced plastome of *R. salina* and *G. theta*.

Even thought there has been nothing to know about the full-detail mechanisms for regulation of *ycf*26 gene expression, an attempt to search for sites involve to gene regulation such as SD sequences, - 35 boxes and -10 boxes that were usually found in the upstream part of functional genes was done. Unfortunately, nothing was detected. Unexpected, the relics of trans-membrane regions were found in front of current *ycf*26 of CCAC 0031.

The phylogenetic tree displayed a very high evolutionary rate of non-Cyanobacterium *ycf*26 in both protein and nucleotide levels in which the cryptophyte clade showed the longest branch.

From these evidences, it is possible to assume that the truncated-*ycf*26 genes without trans-membranes, sensor and HAMP domains in Cryptophyta as well as Raphidophyta lost their functions and on the way to become pseudogenes. Moreover, the degradation pathway seemed to be favorite starting at the N-terminus rather than at C-terminus; it means they lost the trans-membrane, the sensor domains at the N-terminus then continued with HATPase_c at the C-terminus. In the other word, the mutation pressure in *ycf*26 gene allowed the appearing of start codon to be easier than stop codon.

Duplessis *et al*., (2007) also implied that the *ycf*26 protein that lacked trans-membrane regions would be a soluble protein and most likely present in the stroma. To verify the location of this protein in the cell, the authors reported that studies were underway using a Tsg1 (*ycf*26) peptide antibody. However, this assumption seemed not to be matched with observation concerning the domain distribution of algae *ycf*26 analyzed in this study.

4.6 The evolutionary pathway of *chl*I seems to difference with those of SSU, LSU, *rps*4 and *rbc*L

Based on the nucleotide and amino acid sequences of *rps*4 and *chl*I genes, the phylogenetic trees of cryptophytes were displayed. However, the obvious differences in substitution rates among the differences clades and genes examined were observed.

Unlike the unacceptable protein tree of *rbc*L amino acid sequences which failed to recover specific clades that were solved in other DNA sequences data sets (Hoef-Emden *et al.*, 2005), the protein sequence of *rps*4 gene revealed an acceptable phylogenic tree in scheme of the branching pattern of terminal clades in comparison to nucleotide trees obtained from itself or with other DNA trees (Hoef-Emden *et al.*, 2005; Hoef-Emden, 2008). Fortunately, the *rps*4 sequences of the second colorless lineage (presented by M1634) was read; thus, the position of this colorless strain in the *rps*4 phylogenetic trees was observed and confronted with other previously results (Hoef-Emden *et al.*, 2005; Hoef-Emden, 2008). This result again confirmed that two colorless lineages differenced in their evolutionary history.

The phylogenetic trees that were reconstructed from both nucleotide and protein *chl*I data sets, however, resolved the positions of the *Cryptomonas* strains in a total different pattern in comparison with all other results. The LB clade, that was formed by the colorless group (M2452, CCAC 0056 and CCAP 977/2a) with other close relative autotrophic *C. lundii* CCAC 0107, *C. gyropyrenoidosa* CCAC 0108, *C. borealis* CCAC 0113 named by their accelerated substitution rates and terminally diverging of colorless lineage, was broken up in *chl*I trees. Moreover, the heterotrophic strains now located at the basally tips without close relatives with extremely high evolutionary rates. The calculated genetic distances pointed out dominantly accelerated evolutionary rates of *chl*I in comparison with those of *rps*4 and *rbc*L. Thus, it was the reason resulted in unusual tree topologies. Due to the lack of M1634 *chl*I sequence, the position of the second colorless was not archived in this study.

These observations suggested that *chl*I gene increased their substitution rate earlier than *rps*4 and *rbc*L genes as well as the elevated evolutionary rates was ordered by *chl*I > *rps*4 > *rbc*L. The other thing was noted that the different evolutionary way of *chl*I gene in comparison with other genes.

Although having moderate size (609 bp), *rps*4 has an evolution rate neither as high as in *chl*I gene nor as low as in *rbc*L gene, producing acceptable phylogenetic trees for both nucleotide and protein levels. Therefore, it seems to be more suitable protein-enoding plastid gene maker for phylogenies than its sisters, *chl*I and *rbc*L genes.

4.7 The colorless group (M2452, CCAC0056, CCAP 977/2a) diverged early or late depended on the gene surveyed.

As mentioned in the previously results, the colorless strains always occupied at the terminally position regardless of gene markers chosen in nucleus, nucleomorph or plastid. However, in the *chl*I phylogenies, its position was changed significantly to basally tip. These results suggested that at least one gene group in the leucoplast genome represented by *chl*I had increased rapidly its rate at the very early of *Cryptomonas* evolutionary history while the other gene included *rps*4, *rbc*L, SSU and LSU just speeded in recently. The *chl*I tree also implied that the colorless group spilt out of its pigmented ancestor in the early stage of *Cryptomonas* evolutionary history while the other surveyed gene supported the colorless group just diverged in recently.

4.8 Hypothetical scenario for the evolutionary history of colorless *Cryptomonas*

In previously study, Hoef-Emden *et al.*, (2005) released several potential caused for lineage-specific rates in correlating with the loss of photosynthesis in the colorless *Cryptomonas*. This study supported some new evidence.

Some genes directly relative to photosynthetic system (for example, *chl*I gene of three-subunit enzyme Mg-chelatase of chlorophyll biosynthesis) had increased their substitution rate at the very early stage of *Cryptomonas* evolutionary history. The elevation, especially, was extremely strong in the colorless strains with unknown reasons resulted in the less or loss of functional constraints on gene products. The photosynthetic process, in compelling conditions, was affected dramatically so that they became inactive. The loss of photosynthesis was the premise of emergence of at least three independently colorless lineages in genus *Cryptomonas*. The successful changing from autotrophic to heterotrophic lifestyle became the second premise for reducing size of plastid genomes in which many genes have been lost or dramatically altered; *rbc*L gene was a notorious example.

4.9 The shift from NNC to NNT in two-fold degenerate NNY codon seems not relative to the relaxation of the functional constraints of plastome protein-coding genes.

In the previously analyzes (Hoef-Emden *et al.*, 2005), the ratio of NNC/NNU in two-fold degenerate NNY codon calculated for *rbc*L gene showed that this ratio was below 1 in the colorless strains and some close pigmented *Cryptomonas* strains while this value was always above 1 in other *Cryptomonas*. The functional constraints of *rbc*L gene in colorless *Cryptomonas*, therefore, was assigned to have been relaxing as well as the gene expression level was on the way to reduce.

Continuing with this analysis, the numbers of NNC and NNU again were counted for *chl*I, *rps*4 genes of all examined cryptophytes and several newly sequenced *rbc*L gene of autotrophic non-*Cryptomonas* strains. The results in the Appendix 21 showed that most ratios of NNC/NNU in *rbc*L gene of non-*Cryptomonas* strains were higher one but this ratio dropped down below one in cases of GAC/GAU codon (aspartate).

Surprisingly, the NNU codons were always preferred over the NNC codon in *rps*4 and *chl*I genes in all cryptophytes regardless of photo- or nonphotosynthetic strains (Table 3.9.2 and Table 3.10.2). From these results, however, the notion of relaxed functional constraints could not be revealed for *rps*4 and *chl*I concerning to their functions in the biological activities of the cells.

Being considered as a well-conserved protein in prokaryotes and eukaryotes, which suggests strong functional constraints on structural evolution, *rps*4 gene was used widely for evolutionary reconstruction in many phylogenies. With the essential role in protein biosynthetic system in the cells, *rps*4 gene with other genes coding for components of the plastid translational apparatus, therefore, have been preferentially retained in some the reduced leukoplast genome such as *Prototheca, Astila*, … Thus, it was hard to make a linking between the shift from NNC to NNU in two-fold degenerate NNY codon in *rps*4 gene with its decreasing selective constraints and also expression level in cryptophyte plastid. Similar explanation was applied for *chl*I gene.

4.10 The usage of NNU codons over the NNC in two-fold degenerate NNY codon seems to be controlled by neutral mutation pressure rather than by selection followed by the gradually acceleration of evolutionary rate.

As mentioned above, the order of accelerated evolutionary rate can be followed *rbc*L, *rps*4 and *chl*I in which those of *rps*4 and *chl*I were dominantly higher than *rbc*L. Obviously, the shift from NNC to NNU did not emerge in most autotrophic *Cryptomonas* strains where the evolutionary rate of *rbc*L was lowest. The NNU increased the numbers over the numbers of NNC in colorless strains in LB clade and some close relative strains in which their substitution rates were higher. The using of NNU spread in all strains of *Cryptomonas* (and non-*Cryptomonas*) in case of *rps*4 and *chl*I of which their rates were surpassing. Therefore, the shift from NNC to NNU codon in the two-fold degenerate NNY codon seemed to be controlled by neutral mutation pressure in order to increase the AT content rather than by selection; and it was followed by the gradient of accelerated evolutionary rate. The other thing was that the codon GAC/GAU could be considered at the first victim of this changing pressure.

Besides the shift from NNC to NNU was recognized, the usage of YNN codons over the RNN codons also was found in *chl*I gene of colorless *Cryptomonas*. The acceleration of YNN codons meant increasing the appearing probability of stop codons (UAG, UAA and UGA) somewhere in the *chl*I gene.

5. PERSPECTIVES

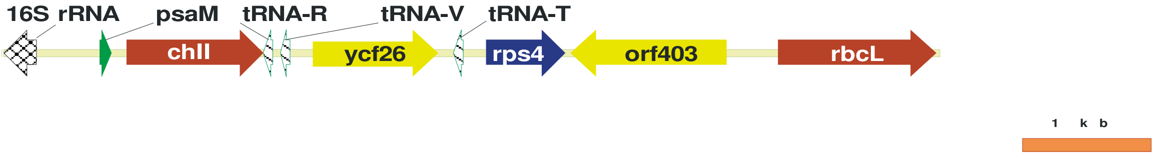
The readings for full-length genomes are always exciting projects. Nowadays, the improvements in strategies and DNA sequencing technology have assisted the genome projects to be done for more efficiently, quickly, costly, *etc*. This study, sequencing fragments of cryptophyte plastomes by combination of long-range PCR and primer-walking sequencing strategies, revealed the basic results of a long journey to discovery full-length cryptophyte plastomes. It is reliable in the future not only to get the sequence database of the same fragments from another cryptophyte strains but also to be able to extent for whole plastome *Cryptomonas* examined. The approach actually opens an ability to harvest huge amount of inter-species plastid genome information quickly and economically. As mentioned in the previously part, the three colorless lineages should be the first chosen for complete leucoplast genome reading; strain *C. erosa* CCAC 0018 and/or *C. obovoidea* CCAC 0031 also to be paid more attention as they are able to possess the biggest platomes among those of *Cryptomonas*. Some plastomes in which their size at the moderate level such as CCAC 0064 also are interesting objects for next stages of the project.

Borza *et al.* (2005) demonstrated that the *chl*I was still expressed after removed to nucleus in colorless *Prototheca wickerhamii*. The authors assumed that the *chl*I gene product was still functional in the communicating regulation between nucleus and plastid genomes. In the heterotrophic *Cryptomonas*, the *chl*I gene product, obviously, has not been participated in the chlorophyll biosynthetic system; however, whether it has been still employed to form the Mg-chelates that involves in the regulation for plastid-to-nucleus retrograde signals? As pointed out, the shift from NNC to NNU in two-fold degenerate NNY codons of *chl*I gene were happened in all examined cryptophytes, the predicting for *chl*I function in colorless strains, therefore, cannot be discussed. Several additive evidences such as the relaxation in 140 last amino acid residues as well as divergence strongly in the second variant region (Fig. 3.9.2) or/and the increased numbers of YNN codon (Table 3.9.2) means increasing the probability of stop codon appearing in colorless *chl*I can be considered as the signals to suggest that the *chl*I gene in colorless to be relaxed the functional constraints and/or on the way to be pseudo-gene. More works and knowledge in the future, however, are needed to get the suitable answer for the relationship in between the present structure and function(s) of *chl*I protein in heterotrophic *Cryptomonas* lineages. Besides that, the same questions can be questioned also to *ycf*26 and ORF403.

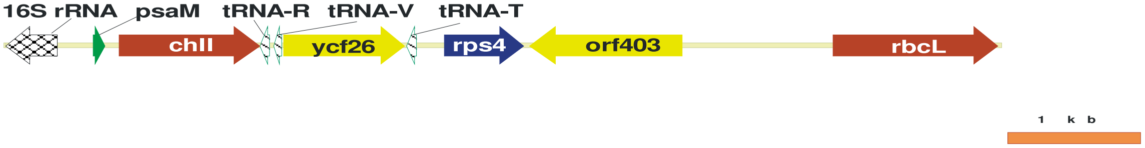
The evolutionary history of colorless *Cryptomonas* is one of the main objects that have been attracting the attention of many molecular systematic investigators. The potential causing of the forming of these lineages as well as the relationship between loss photosynthesis and increased substitution rates were published recently. This study, also, discussed one of hypothetical scenario to explain the extremely acceleration of photosynthetic genes with correlating with the changing from autotrophic to heterotrophic lifestyle of colorless lineages turning to be the main factor affects the increasing evolutionary rate of other genes. As any new hypothesis, it will be accepted or rejected if new lines of evidence will be available in the future.

6. APPENDIX

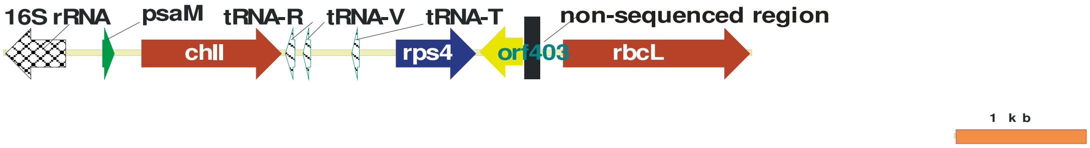
Appendix 1: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Rhodomonas* sp. M1480 plastome



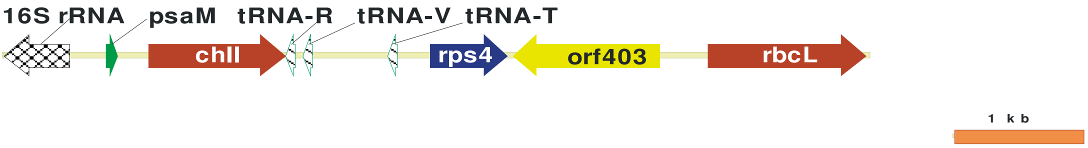
Appendix 2: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Teleaulax* sp. SCAP K-416 plastome



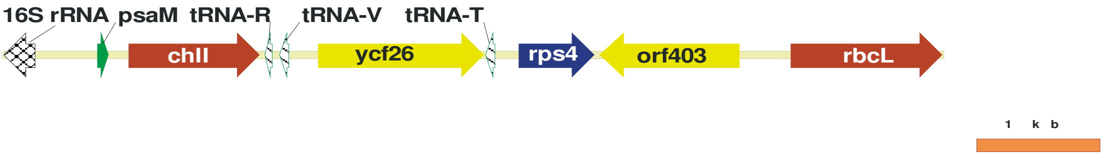
Appendix 3: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. commutata* CCAC 0109 plastome



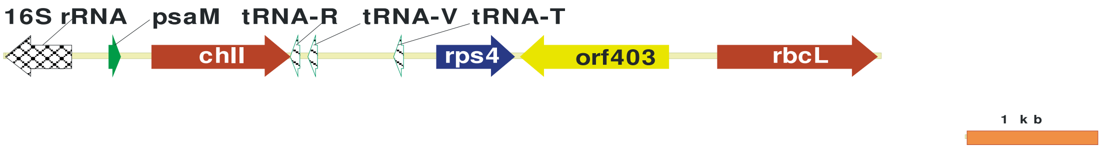
Appendix 4: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. loricata* M 2088 plastome



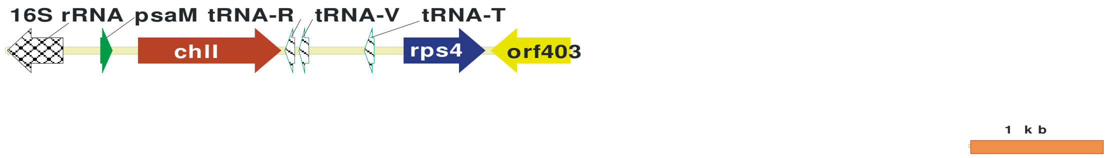
Appendix 5: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. erosa* CCAC 0018 plastome



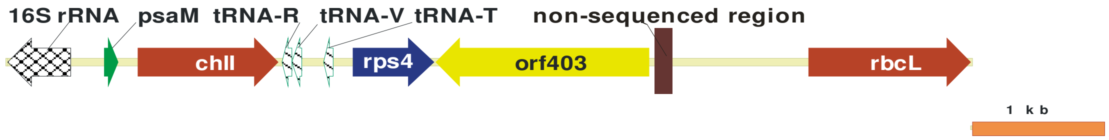
Appendix 6: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. loricata* M 2088 plastome



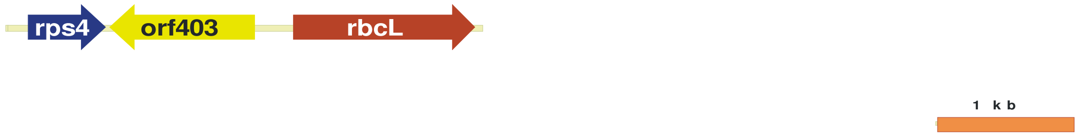
Appendix 7: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. curvata* CCAC 0006 plastome



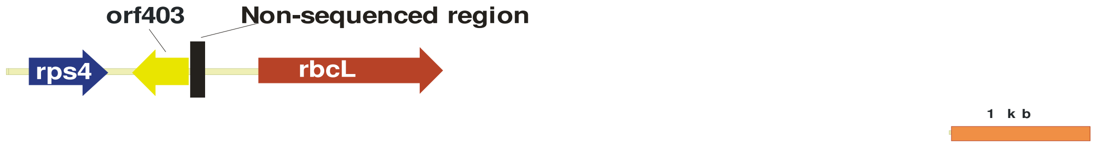
Appendix 8: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. borealis* CCAC 0113 plastome



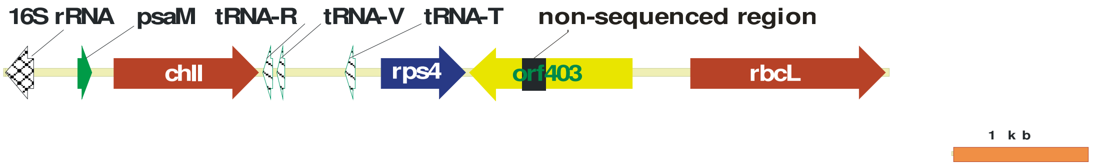
Appendix 9: Illustration the gene content, gene order of the *rps*4 – *rbc*L fragment in *C. gyropyrenoidosa* sp. *nov*. CCAC 0108 plastome



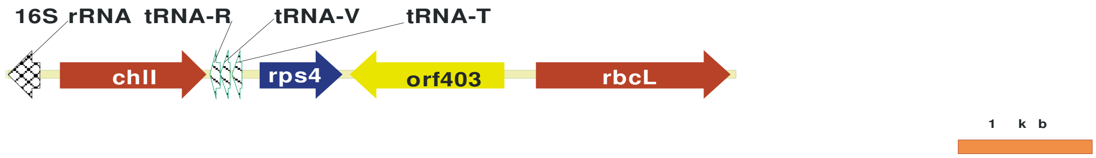
Appendix 10: Illustration the gene content, gene order of the *rps*4–*rbc*L fragment in *C. lundii* CCAC 0107 plastome



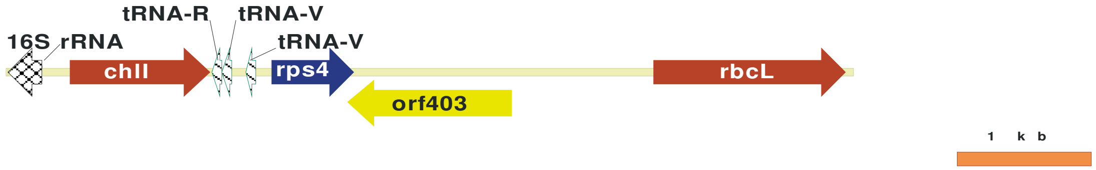
Appendix 11: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. ovata* CCAC 0064 plastome



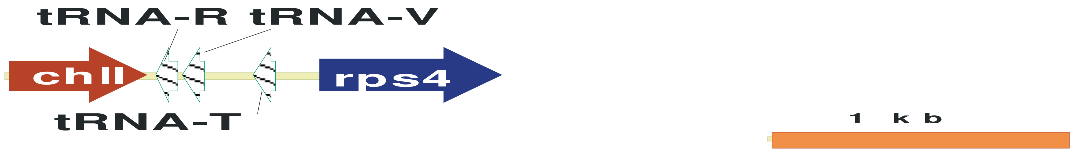
Appendix 12: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *C. paramaecium* CCAP 977/2a plastome



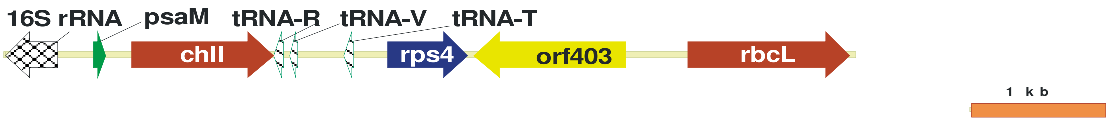
Appendix 13: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. paramaecium* CCAC 0056 plastome



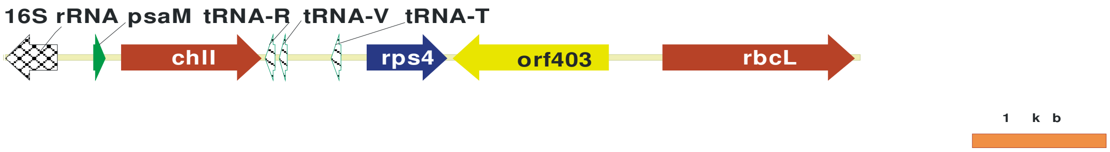
Appendix 14: Illustration the gene content, gene order of the *chl*I–*rps*4 fragment in *C.* M1634 plastome



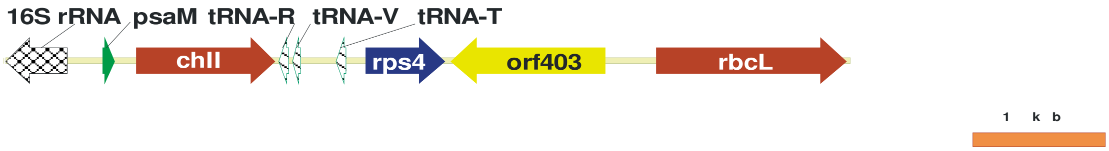
Appendix 15: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment *in C. phaseolus* SAG 2013 plastome



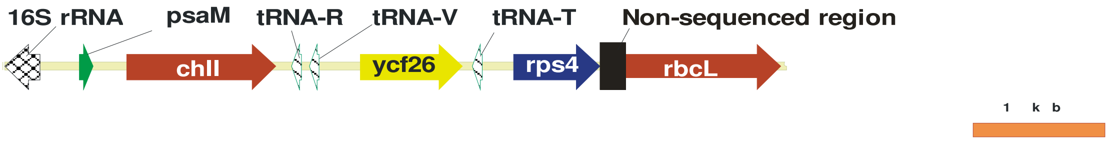
Appendix 16: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. pyrenoidifera* CCMP 0152 plastome



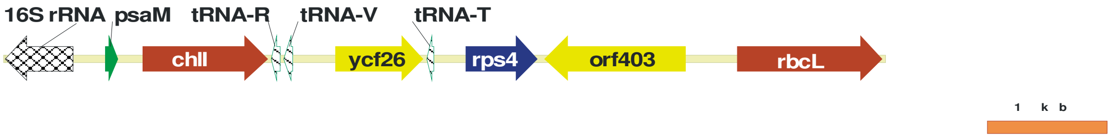
Appendix 17: Illustration the gene content, gene order of the 16S rRNA–*rbc*L fragment in *C. tetrapyrenoidosa* M 1092plastome



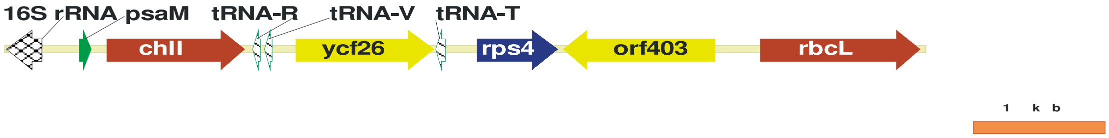
Appendix 18: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Chroomonas* sp. SAG 980-1 plastome



Appendix 19: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Hemiselmis tepida* CCMP 0443 plastome



Appendix 20: Illustration the gene contents, gene orders of the 16S rRNA–*rbc*L fragment in *Proteomonas* sp. CCMP 0704 plastome



Appendix 21: NNC and NNU codon numbers in *rbc*L gene of some non-*Cryptomonas* strains

	STRAINS	AAC/AAU ASN	CAC/CAU HIS	GAC/GAU ASP	UAC/UAU TYR	UGC/UGU CYS	UUC/UUU PHE
1.	SAG 2013	12/1	8/1	9/8	11/4	1/5	14/2
2.	SAG 980-1	12/1	5/4	7/10	6/10	0/5	10/5
3.	CCMP 0443	14/0	7/2	6/10	16/10	0./7	14/1
4.	CCPM 0704	15/0	9/0	7/8	13/0	1/4	17/1
5.	M1480	11/2	7/1	8/10	11/4	1/4	15/1
6.	SCCAP K-416	7/6	2/7	5/11	7/8	1/2	8/8
7.	CCMP 1319	11/2	7/1	8/10	11/4	1/4	15/1

Appendix 22: Accession numbers of Cyanobateria *ycf*26; AT% of the cyanobacterial genomes and *ycf*26 genes

	STRAIN	ACCESSION NUMBER	Length of ycf26	AT % of ycf26	AT% of Cyanomes
1.	***Acaryochloris marina*** **MBIC11017**	**NC_009925**	**1917**	**51**	**53**
2.	***Anabaena variabilis*** **ATCC 29413**	**NC_007413**	**1953**	**55**	**59**
3.	***Crocosphaera watsonii*** **WH 8501**	**ZP_00518795**	**2040**	**64**	**63**
4.	***Cyanothece*** **sp. PCC 8801**	**ZP_02941120**	**1968**	**60**	**60**
5.	***Cyanothece*** **sp. CCY0110**	**ZP_01731146**	**2037**	**62**	**63**
6.	***Cynechococcus*** **sp. RS9916**	**NZ_AAUA01000001**	**2052**	**38**	**40**
7.	***Gloeobacter violaceus*** **PCC 7421**	**NP_923001**	**1935**	**39**	**48**
8.	***Lyngbya*** **sp. PCC 8106**	**ZP_01618936**	**2256**	**59**	**59**
9.	***Microcystis aeruginosa*** **NIES843**	**NC_010296**	**1998**	**59**	**58**
10.	***Microcystis aeruginosa*** **PCC 7806**	**CAO89196**	**1998**	**59**	**59**
11.	***Nodularia spumigena*** **CCY9414**	**ZP_01628860**	**1947**	**57**	**59**
12.	***Nostoc punctiforme*** **PCC 73102**	**NZ_AAAY02000005**	**1947**	**56**	**59**
13.	***Nostoc punctiforme*** **PCC 73102**	**NC_010628**	**1947**	**56**	**59**
14.	***Nostoc*** **sp. PCC 7120**	**NC_003272**	**1860**	**59**	**41**
15.	***Prochlorococcus marinus*** **clone HF1088H9**	**DQ366729**	**2067**	**56**	**68**
16.	***Prochlorococcus marinus*** **clone HOT0M1A11**	**DQ366734**	**2067**	**65**	**68**
17.	***Prochlorococcus marinus*** **str. MIT 9515**	**NC_008817**	**2070**	**67**	**69**
18.	***Prochlorococcus marinus*** **str. MIT 9211**	**NC_009976**	**2061**	**63**	**62**
19.	***Prochlorococcus marinus*** **str. MIT 9215**	**NC_009840**	**2067**	**65**	**69**
20.	***Prochlorococcus marinus*** **str. MIT 9301**	**NC_009091**	**2067**	**65**	**69**
21.	***Prochlorococcus marinus*** **str. MIT 9303**	**CP000554**	**2058**	**47**	**50**
22.	***Prochlorococcus marinus*** **str. MIT 9312**	**NC_007577**	**2070**	**66**	**69**

23.	*Prochlorococcus marinus* str. MIT 9313	NC_005071	2058	48	49
24.	*Prochlorococcus marinus* str. NATL1A	NC_008819	2070	64	65
25.	*Prochlorococcus marinus* str. NATL2A	NC_007335	2070	64	65
26.	*Prochlorococcus marinus* subsp. *marinus* str. CCMP1375	NC_005042	2070	64	64
27.	*Prochlorococcus marinus* subsp. *pastoris* str. CCMP1986	NC_005072	2070	67	69
28.	*Synechococcus elongatus* PCC 6301	NC_006576	1887	45	44
29.	*Synechococcus elongatus* PCC 7942	NC_007604	1992	45	47
30.	*Synechococcus* sp. BL107	NZ_AATZ01000001	2070	43	46
31.	*Synechococcus* sp. CC9311	NC_008319	2082	46	48
32.	*Synechococcus* sp. CC9605	NC_007516	2067	37	41
33.	*Synechococcus* sp. CC9902	NC_007513	2091	43	46
34.	*Synechococcus* sp. JA23B'a(213)	NC_007776	2103	40	41
35.	*Synechococcus* sp. JA33Ab	NC_007775	2004	38	40
36.	*Synechococcus* sp. PCC 7002	NC_010475	2019	51	50
37.	*Synechococcus* sp. RCC307	NC_009482	1935	38	39
38.	*Synechococcus* sp. RS9917	NZ_AANP01000002	2058	38	35
39.	*Synechococcus* sp. WH 5701	NZ_AANO01000002	2004	36	35
40.	*Synechococcus* sp. WH 7803	NC_009481	2046	39	40
41.	*Synechococcus* sp. WH 7805	NZ_AAOK01000001	2052	41	42
42.	*Synechococcus* sp. WH 8102	NC_005070	2040	37	41
43.	*Synechocystis* sp. PCC 6803	BA000022	1918	52	56
44.	*Thermosynechococcus elongatus* BP-1	NC_004113	1956	49	54
45.	*Trichodesmium erythraeum* IMS101	NC_008312	2148	66	66

7. REFERENCES

Abascal, F., Zardoya, R. & Posada, D. (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**(9), 2104-2105.

Adachi, J., P. Waddell, W. Martin, and M. Hasegawa (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution* **50**, 348-358.

Andersson, S.G. & Kurland, C.G. (1990) Codon preferences in free-living microorganisms. *Microbiology and Molecular Biology Reviews* **54**(2), 198-210.

Ashby, M.K., Houmard, J. & Mullineaux, C.W. (2002) The *ycf*27 genes from cyanobacteria and eukaryotic algae: distribution and implications for chloroplast evolution. *FEMS Microbiology Letters* **214**(1), 25-30.

Ashby, M.K. & Houmard, J. (2006) Cyanobacterial two-component proteins: structure, diversity, distribution, and evolution. *Microbiology and Molecular Biology Reviews* **70**(2), 472-509.

Bollivar, D. (2006) Recent advances in chlorophyll biosynthesis. *Photosynthesis Research* **90**, 173-194.

Borza, T., Popescu, C.E. & Lee, R.W. (2005) Multiple metabolic roles for the nonphotosynthetic plastid of the green alga *Prototheca wickerhamii*. *Eukaryotic Cell* **4**(2), 253-261.

Cavalier-Smith, T., Couch, J.A., Thorsteinsen, K.E., Gilson, P., Deane, J.A., Hill, D.R.A. & Mcfadden, G.I. (1996) Cryptomonad nuclear and nucleomorph 18S rRNA phylogeny. *European Journal of Phycology* **31**, 315-328.

Cavalier-Smith, T. (2002) Nucleomorphs: enslaved algal nuclei. *Current Opinion in Microbiology* **5**, 612-619.

Cheng, S., Chang, S.Y., Gravitt, P. & Respess, R. (1994) Long PCR. *Nature* **369**(6482), 684-685.

Clay, B.L., Kugrens, P. & Lee, R.E. (1999) A revised classification of Cryptophyta. *Botanical Journal of the Linnean Society* **131**, 131-151.

Clay, B.L., Paul, K. & Robert, E.L. (2001) Cryptomonads. In *Encyclopedia of Life Sciences*: John Wiley & Sons, Ltd.

Davies, C., Gerstner, R.B., Draper, D.E., Ramakrishnan, V. & White, S.W. (1998) The crystal structure of ribosomal protein S4 reveals a two-domain molecule with an extensive RNA-binding surface: one domain shows structural homology to the ETS DNA-binding motif. *The EMBO Journal* **17**(16), 4545-4558.

Deane, J.A., Strachan, I.M., Saunders, G.W., Hill, D.R.A. & McFadden, G.I. (2002) Cryptomonad evolution: nuclear 18S rDNA phylogeny versus cell morphology and pigmentation. *Journal of Phycology* **38**(6), 1236-1244.

Douglas, S. (1988) Physical mapping of the plastid genome from the chlorophyll c-containing alga, *Cryptomonas* Φ. *Current Genetics* **14**, 591-598.

Douglas, S.E. & Durnford, D.G. (1990) Nucleotide sequence of the genes for ribosomal protein S4 and tRNA-Arg from the chlorophyll c-containing alga *Cryptomonas* □. *Nucleic Acids Reseach* **18**(7), 1903.

Douglas, S.E., Durnford, D.G. & Morden, C.W. (1990) Nucleotide sequence of the gene for the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from *Cryptomonas* Φ : Evidence supporting the polyphyletic origin of plastids. *Journal of Phycology* **26**(3), 500-508.

Douglas, S.E. & Murphy, C.A. (1994) Structural, transcriptional, and phylogenetic analyses of the atpB gene cluster from the plastid of *Cryptomonas* (Cryptophyceae). *Journal of Phycology* **30**(2), 329-340.

Douglas, S.E. & Penny, S.L. (1999) The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. *Journal of Molecular Evolution* **48**(2), 236-244.

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.-T., Wu, X., Reith, M., Cavalier-Smith, T. & Maier, U.-G. (2001) The highly reduced genome of an enslaved algal nucleus. *Nature* **410**(6832), 1091-1096.

Doyle, J. & Doyle, D.J. (1990) Isolation of plant DNA from fresh tissue. *Focus* **12:**, 13-15.

Duplessis, M.R., Karol, K.G., Adman, E.T., Choi, L.Y., Jacobs, M.A. & Cattolico, R.A. (2007) Chloroplast His-to-Asp signal transduction: a potential mechanism for plastid gene regulation in *Heterosigma akashiwo* (Raphidophyceae). *BMC Evolutionary Biology* **7**, 70.

Fodje, M.N., Hansson, A., Hansson, M., Olsen, J.G., Gough, S., Willows, R.D. & Al-Karadaghi, S. (2001) Interplay between an AAA module and an integrin I domain may regulate the function of magnesium chelatase. *Journal of Molecular Biology* **311**(1), 111-122.

Fromme, P., Jordan, P. & Krauss, N. (2001) Structure of photosystem I. *Biochimica et Biophysica Acta* **1507**, 5-31.

Galtier, N., Gouy, M. & Gautier, C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics* **12**(6), 543-548.

Giesecke, H., Obermaier, B., Domdey, H. & Neubert, W.J. (1992) Rapid sequencing of the Sendai virus 6.8 kb large (L) gene through primer walking with an automated DNA sequencer. *Journal of Virological Methods* **38**(1), 47-60.

Goremykin, V., Hirsch-Ernst, K.I., Wölfl, S. & Hellwig, F.H. (2003a) The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis* – structural and phylogenetic analyses. *Plant Systematics and Evolution* **242**(1), 119-135.

Goremykin, V.V., Hirsch-Ernst, K.I., Wolfl, S. & Hellwig, F.H. (2003b) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that amborella is not a basal angiosperm. *Molecular Biology and Evolution* **20**(9), 1499-1505.

Goremykin, V.V., Hirsch-Ernst, K.I., Wolfl, S. & Hellwig, F.H. (2004) The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. *Molecular Biology and Evolution* **21**(7), 1445-1454.

Gould, S., Sommer, M., Hadfi, K., Zauner, S., Kroth, P. & Maier, U.-G. (2006) Protein targeting into the complex plastid of cryptophytes. *Journal of Molecular Evolution* **62**(6), 674-681.

Gould, S.B., Sommer, M.S., Kroth, P.G., Gile, G.H., Keeling, P.J. & Maier, U.-G. (2006) Nucleus-to-nucleus gene transfer and protein retargeting into a remnant cytoplasm of cryptophytes and diatoms. *Molecular Biology and Evolution* **23**(12), 2413-2422.

Gray, J.C. (2003) Chloroplast-to-nucleus signalling: a role for Mg-protoporphyrin. *Trends in Genetics* **19**(10), 526-529.

Grzebyk, D., Schofield, O., Vetriani, C. & Falkowski, P.G. (2003) The mesozoic radiation of eukaryotic algae: The portable plastid hypothesis. *Journal of Phycology* **39**(2), 259-267.

Guindon, S. & Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**(5), 696-704.

Harris, E.H., Boynton, J.E. & Gillham, N.W. (1994) Chloroplast ribosomes and protein synthesis. *Microbiology and Molecular Biology Reviews* **58**(4), 700-754.

Hauth, A.M., Maier, U.G., Lang, B.F. & Burger, G. (2005) The *Rhodomonas salina* mitochondrial genome: bacteria-like operons, compact gene arrangement and complex repeat region. *Nucleic Acids Research* **33**(14), 4433-4442.

Hillis, D.M., Dixon, M.T. (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly Review of Biology* **66**(4), 411-453.

Hirose, T. & Sugiura, M. (2004) Functional Shine-Dalgarno-like sequences for translational initiation of chloroplast mRNAs. *Plant & Cell Physiology* **45**(1), 114-117.

Hoef-Emden, K., Marin, B. & Melkonian, M. (2002) Nuclear and nucleomorph SSU rDNA phylogeny in the Cryptophyta and the evolution of cryptophyte diversity. *Journal of Molecular Evolution.* **55**, 161-179.

Hoef-Emden, K. & Melkonian, M. (2003) Revision of the genus *Cryptomonas* (Cryptophyceae): A combination of molecular phylogeny and morphology provides insights into a long-hidden dimorphism. *Protist* **154**, 371-409.

Hoef-Emden, K., Tran, H.-D. & Melkonian, M. (2005) Lineage-specific variations of congruent evolution among DNA sequences from three genomes, and relaxed selective constraints on *rbc*L in *Cryptomonas* (Cryptophyceae). *BMC Evolutionary Biology* **5**(56).

Hoef-Emden, K. (2005a) Multiple independent losses of photosynthesis in the genus *Cryptomonas* (Cryptophyceae) - combined phylogenetic analyses of DNA sequences of the nuclear and the nucleomorph ribosomal operons. *Journal of Molecular Biology* **60**, 183-195.

Hoef-Emden, K. (2005b) Molecular phylogenetic analyses and real-life data. *Computing in Science and Engineering* **7**(3), 86-91.

Hoef-Emden, K. (2007) Revision of the genus *Cryptomonas* (Cryptophyceae) II: Incongruences between the classical morphospecies concept and molecular phylogeny in smaller pyrenoid-less cells. *Phycologia* **46**(4), 402-428.

Hoef-Emden, K. & Archibald., J.M. (2008a) Cryptomonads = *Goniomonas* + plastid-containing cryptophytes. Cryptophyta. Version 14 September 2008. *http://tolweb.org/Cryptomonads/2396/2008.09.14* in The Tree of Life Web Project, http://tolweb.org/.

Hoef-Emden, K. (2008b) Molecular phylogeny of phycocyanin-containing cryptophytes: evolution of biliproteins and geographical distribution. *Journal of Phycology* **44**(4), 985-993.

Hu, M., Chilton, N.B. & Gasser, R.B. (2002) Long PCR-based amplification of the entire mitochondrial genome from single parasitic nematodes. *Molecular and Cellular Probes* **16**(4), 261-267.

Jansen, R.K., Raubeson, L.A., Boore, J.L., dePamphilis, C.W., Chumley, T.W., Haberle, R.C., Wyman, S.K., Alverson, A.J., Peery, R., Herman, S.J., Fourcade, H.M., Kuehl, J.V., McNeal, J.R., Leebens-Mack, J. & Cui, L. (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology* **395**, 348-384.

Jensen, P.E., Gibson, L.C.D., Henningsen, K.W. & Hunter, C.N. (1996) Expression of the *chl*I, *chl*D, and *chl*H genes from the cyanobacterium *Synechocystis* PCC6803 in *Escherichia coli* and demonstration that the three cognate proteins are required for magnesium-protoporphyrin chelatase activity. *Journal of Biological Chemsitry* **271**(28), 16662-16667.

Kalanon, M. & McFadden, G.I. (2008) The chloroplast protein translocation complexes of *Chlamydomonas reinhardtii*: a bioinformatic comparison of Toc and Tic components in plants, green algae and red algae. *Genetics* **179**(1), 95-112.

Kanesaki, Y., Yamamoto, H., Paithoonrangsarid, K., Shoumskaya, M., Suzuki, I., Hayashi, H. & Murata, N. (2007) Histidine kinases play important roles in the perception and signal transduction of hydrogen peroxide in the cyanobacterium, *Synechocystis* sp. PCC 6803. *The Plant journal* **49**(2), 313-324.

Kellogg, E.A. & Juliano, N.D. (1997) The structure and function of RuBisCO and their implications for systematic studies. *American Journal of Botany* **84**(3), 413-428.

Khan, H., Kozera, C., Curtis, B., Bussey, J., Theophilou, S., Bowman, S. & Archibald, J. (2007a) Retrotransposons and tandem repeat sequences in the nuclear genomes of cryptomonad algae. *Journal of Molecular Evolution* **64**(2), 223-236.

Khan, H., Parks, N., Kozera, C., Curtis, B.A., Parsons, B.J., Bowman, S. & Archibald, J.M. (2007b) Plastid genome sequence of the cryptophyte alga *Rhodomonas salina* CCMP1319: lateral transfer of putative DNA replication machinery and a test of chromist plastid phylogeny. *Molecular Biology and Evolution* **24**(8), 1832-1842.

Kies L (1967) Über Zweiteilung und Zygotenbildung bei Roya obtusa (Bre.) West & West. *Mitt Staatsint Allg Bot Hamb* **12**:35-42

Kim, I.C., Jung, S.O., Lee, Y.M., Lee, C.J., Park, J.K. & Lee, J.S. (2005) The complete mitochondrial genome of the rayfish *Raja porosa* (Chondrichthyes, Rajidae). *DNA Sequence* **16**(3), 187-194.

Kim, E., Lane, C.E., Curtis, B.A., Kozera, C., Bowman, S. & Archibald, J.M. (2008) Complete sequence and analysis of the mitochondrial genome of *Hemiselmis andersenii* CCMP644 (Cryptophyceae). *BMC Genomics.* **9**(245).

Kouranov, A. & Schnell, D.J. (1997) Analysis of the Interactions of preproteins with the import machinery over the course of protein Import into chloroplasts. *The Journal of Cell Biology* **139**(7), 1677-1685.

Kouranov, A., Chen, X., Fuks, B. & Schnell, D.J. (1998) Tic20 and Tic22 are new components of the protein import apparatus at the chloroplast inner envelope membrane. *The Journal of Cell Biology* **143**(4), 991-1002.

Kumar, S., Dudley, J., Nei, M. & Tamura, K. (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics* **9**, 299-306.

Lane, C.E. & Archibald, J.M. (2006a) Novel nucleomorph genome architecture in the cryptomonad genus *Hemiselmis*. *The Journal of Eukaryotic Microbiology* **53**(6), 515-521.

Lane, C.E., Khan, H., Fong, A., Theophilou, S. & Archibald, J.M. (2006b) Insight into the diversity and evolution of the cryptomonad nucleomorph genome. *Molecular Biology and Evolution* **23**, 856-865.

Lee, R.E. (1999) Cryptophyta. In *Phycology (third editions)*, pp. 365-377. Cambridge University Press.

Lerat, E. & Ochman, H. (2005) Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Research* **33**(10), 3125-3132.

Maerz, M., Wolters, J., Hofmann, C.J., Sitte, P. & Maier, U.G. (1992) Plastid DNA from *Pyrenomonas salina* (Cryptophyceae): physical map, genes, and evolutionary implications. *Current Genetics* **21**(1), 73-81.

Maier, U.G. (1992) The four genomes of the alga *Pyrenomonas salina* (Cryptophyta). *Biosystems* **28**(1-3), 69-73.

Maier, U.G., Douglas, S.E. & Cavalier-Smith, T. (2000) The nucleomorph genomes of cryptophytes and chlorarachniophytes. *Protist* **151**, 103-109.

Marin, B., Klingberg, M. & Melkonian, M. (1998) Phylogenetic relationships among the Cryptophyta: analyses of nuclear-encoded SSU rRNA sequences support the monophyly of extant plastid-containing lineages. *Protist* **149**, 265-276.

Markus, M.A., Gerstner, R.B., Draper, D.E. & Torchia, D.A. (1998) The solution structure of ribosomal protein S4 delta41 reveals two subdomains and a positively charged surface that may interact with RNA. *The EMBO Journal* **17**(16), 4559-4571.

Martin, W., Stoebe, B., Goremykin, V., Hapsmann, S., Hasegawa, M. & Kowallik, K.V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**(6681), 162-165.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M. & Penny, D. (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences* **99**(19), 12246-12251.

Matsuzaki, M., Misumi, O., Shin-i, T., Maruyama, S., Takahara, M., Miyagishima, S.-y., Mori, T., Nishida, K., Yagisawa, F., Nishida, K., Yoshida, Y., Nishimura, Y., Nakao, S., Kobayashi, T., Momoyama, Y., Higashiyama, T., Minoda, A., Sano, M., Nomoto, H., Oishi, K., Hayashi, H., Ohta, F., Nishizaka, S., Haga, S., Miura, S., Morishita, T., Kabeya, Y., Terasawa, K., Suzuki, Y., Ishii, Y., Asakawa, S., Takano, H., Ohta, N., Kuroiwa, H., Tanaka, K., Shimizu, N., Sugano, S., Sato, N., Nozaki, H., Ogasawara, N., Kohara, Y. & Kuroiwa, T. (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* **428**(6983), 653-657.

McFadden, G. & Melkonian, M. (1986) Use of hepes buffer for microalgal culture media and fixation for electron microscopy. *Phycologia* **25**, 551-557.

McFadden, G.I. & van Dooren, G.G. (2004) Evolution: red algal genome affirms a common origin of all plastids. *Current Biology* **14**(13), R514-516.

Melkonian, M., Marin, B. & Surek, B. (1995) Phylogeny and evolution of algae. In *Biodiversity and evolution*, pp. 153-176. Edited by R. Arai, Kato, M., Doi, Y. Tokyo: The National Science Museum Foundation.

Mereu, P., Suni, M.P.d., Manca, L. & Masala, B. (2007) Complete nucleotide mtDNA sequence of Barbary sheep (*Ammotragus lervia*). *DNA Sequence* **1**.

Mikami, K., Kanesaki, Y., Suzuki, I. & Murata, N. (2002) The histidine kinase Hik33 perceives osmotic stress and cold stress in *Synechocystis* sp PCC 6803. *Molecular Microbiology* **46**(4), 905-915.

Mikami, K., Suzuki, I. & Murata, N. (2003) Sensors of abiotic stress in *Synechocystis*. In *Plant Responses to Abiotic Stress*, pp. 103-119.

Morici, L., Frisk, A. & Schurr, M. (2006) Two-component regulatory systems. In *Molecular Paradigms of Infectious Disease*, pp. 502-543.

Morrison, S.S., Mullineaux, C.W. & Ashby, M.K. (2005) The influence of acetyl phosphate on DspA signalling in the Cyanobacterium *Synechocystis* sp. PCC6803. *BMC Microbiology* **5**, 47.

Munemasa, M., Nikaido, M., Donnellan, S., Austin, C.C., Okada, N. & Hasegawa, M. (2006) Phylogenetic analysis of diprotodontian marsupials based on complete mitochondrial genomes. *Genes & Genetic Systems* **81**(3), 181-191.

Murray, M.G. & Thompson, W.F. (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research* **8**(19), 4321-4326.

Naithani, S., Hou, J.M. & Chitnis, P.R. (2000) Targeted inactivation of the *psa*K1, *psa*K2 and *psa*M genes encoding subunits of Photosystem I in the cyanobacterium *Synechocystis* sp. PCC 6803. *Photosynthesis Research* **63**, 225-236.

Nakayama, M., Masuda, T., Sato, N., Yamagata, H., Bowler, C., Ohta, H., Shioi, Y. & Takamiya, K. (1995) Cloning, subcellular localization and expression of *CHL*1, a subunit of magnesium-chelatase in soybean. *Biochemical and Biophysical Research Communications* **215**(1), 422-428.

Nelson, N. & Yocum, C.F. (2006) Structure and function of photosystems I and II. *Annual Review of Plant Biology* **57**(1), 521-565.

Nott, A., Jung, H.S., Koussevitzky, S. & Chory, J. (2006) Plastid-to-nucleus retrograde signaling. *Annual Review of Plant Biology* **57**, 739-759.

Novarino, G. (2003) A companion to the identification of cryptomonad flagellates (Cryptophyceae = Cryptomonadea). *Hydrobiologia* **502**(1), 225-270.

Nozaki, H., Ohta, N., Matsuzaki, M., Misumi, O. & Kuroiwa, T. (2003) Phylogeny of plastids based on cladistic analysis of gene loss inferred from complete plastid genome sequences. *Journal of Molecular Evolution* **57**(4), 377-382.

Ohta, N., Sato, N., Ueda, K. & Kuroiwa, T. (1997) Analysis of a plastid gene cluster reveals a close relationship between *Cyanidioschyzon* and *Cyanidium*. *Journal of Plant Research* **110**(2), 235-245.

Osawa, S., Jukes, T.H., Watanabe, K. & Muto, A. (1992) Recent evidence for evolution of the genetic code. *Microbiology and Molecular Biology Reviews* **56**(1), 229-264.

Page, R.D.M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**, 357-358.

Palumbi, S.R. (1990) Nucleic acids II: The polymerase chain reaction. In *Molecular Systematics*, pp. 212-214. Edited by D.M. Hillis, C.M. Sunderland & B. Mable. Sinauer Associates Inc.

Phipps, K.D., Donaher, N.A., Lane, C.E. & Archibald, J.M. (2008) Nucleomorph karyotype diversity in the freshwater cryptophyte genus *Cryptomonas*. *Journal of Phycology* **44**(1), 11-14.

Ponce, M.R. & Micol, J.L. (1992) PCR amplification of long DNA fragments. *Nucleic Acids Research* **20**(3), 623.

Posada, D. (2003) Selecting models of evolution. In *The Phylogenetic Handbook*, pp. 256-282. Edited by A.M. Vandemme & M. Salemi. Cambridge University press.

Posada, D. (2008) jModelTest: phylogenetic model averaging. *Molecular biology and evolution* **25**(7), 1253-1256.

Posada, D. & Crandall, K.A. (2001) Selecting the best-fit model of nucleotide substitution. *Systematic Biology* **50**(4), 580-601.

Reith, M. & Munholland, J. (1993) A high-resolution gene map of the chloroplast genome of the red alga *Porphyra purpurea*. *The Plant Cell* **5**(4), 465-475.

Reumann, S., Inoue, K. & Keegstra, K. (2005) Evolution of the general protein import pathway of plastids (review). *Molecular Membrane Biology* **22**(1-2), 73-86.

Rodermel, S. (2001) Pathways of plastid-to-nucleus signaling. *Trends in Plant Science* **6**(10), 471-478.

Ronquist, F. & Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**(12), 1572-1574.

Sambrook, J. & Russell, D.W. (2006) Fluorometric Quantitation of DNA Using Hoechst 33258. *Cold Spring Harbor Protocols* **2006**(1), pdb.prot4458-.

Sato, N. (2006) Origin and evolution of plastids: Genomic view on the unification and diversity of plastids. In *The Structure and Function of Plastids*, pp. 75-102. Springer.

Schubert, W.D., Klukas, O., Krausz, N., Saenger, W., Fromme, P. & Witt, H.T. (1997) Photosystem I of *Synechococcus elongatus* at 4A resolution: comprehensive structure analysis. *Journal of Molecular Biology* **272**, 741-769.

Somanchi, A. & Mayfield, S. (2004) Regulation of chloroplast translation. In *Regulation of Photosynthesis*, pp. 137-151.

Spreitzer, R.J. & Salvucci, M.E. (2002) Rubisco: structure, regulatory interactions, and possibilities for a better enzyme. *Annual Review of Plant Biology* **53**, 449-75.

Sterky, F. & Lundeberg, J. (2000) Sequence analysis of genes and genomes. *Journal of Biotechnology* **76**(1), 1-31.

Stock, A.M., Robinson, V.L. & Goudreau, P.N. (2000) Two-component signal transduction. *Annual Review of Biochemistry* **69**(1), 183-215.

Stoebe, B., Martin, W. & Kowallik, K.V. (1998) Distribution and nomenclature of protein-coding genes in 12 sequenced chloroplast genomes. *Plant Molecular Biology Reporter* **16**(3), 243-255.

Strand, A., Asami, T., Alonso, J., Ecker, J.R. & Chory, J. (2003) Chloroplast to nucleus communication triggered by accumulation of Mg-protoporphyrinIX. *Nature* **421**(6918), 79-83.

Suzuki, I., Los, D.A., Kanesaki, Y., Mikami, K. & Murata, N. (2000) The pathway for perception and transduction of low-temperature signals in *Synechocystis*. *The EMBO Journal* **19**(6), 1327-1334.

Suzuki, J.Y., Bollivar, D.W. & Bauer, C.E. (1997) Genetic analysis of chlorophyll biosynthesis. *Annual Review of Genetics* **31**(1), 61-89.

Swofford, D.L. (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* Sinauer Associates, Sunderland, Massachusetts.

Tanifuji, G., Erata, M., Ishida, K.-i., Onodera, N. & Hara, Y. (2006) Diversity of secondary endosymbiont-derived actin-coding genes in cryptomonads and their evolutionary implications. *Journal of Plant Research* **119**(3), 205-215.

Valentin, K., Fischer, S. & Ann Cattolico, R. (1998) The chloroplast *bchI* gene encodes a subunit of magnesium chelatase in the marine heterokont alga *Heterosigma carterae*. *European Journal of Phycology* **33**, 113-120.

van den Hoek, C., Mann, D.G., Jahns, H.M. (1996) Cryptophyta. In *Algae. An Introduction to Phycology*, pp. 235-243. Cambridge University Press.

van Waasbergen, L.G., Dolganov, N. & Grossman, A.R. (2002) nblS, a gene involved in controlling photosynthesis-related gene expression during high light and nutrient stress in *Synechococcus elongatus* PCC 7942. *The Journal of Bacteriology* **184**(9), 2481-2490.

Waller, R.F. & McFadden, G.I. (2005) The apicoplast: a review of the derived plastid of apicomplexan parasites. *Current Issues in Molecular Biology* **7**(1), 57-79.

Wang, S.L., Liu, X.Q. & Douglas, S.E. (1997) The large ribosomal protein gene cluster of a cryptomonad plastid: gene organization, sequence and evolutionary implications. *Biochem Mol Biol Int* **41**(5), 1035-1044.

Wong, J.T. & Cedergren, R. (1986) Natural selection versus primitive gene structure as determinant of codon usage. *European Journal of Biochemistry* **159**(1), 175-180.

Yamauchi, M.M., Miya, M.U., Machida, R.J. & Nishida, M. (2004a) PCR-based approach for sequencing mitochondrial genomes of decapod crustaceans, with a practical example from kuruma prawn (*Marsupenaeus japonicus*). *Marine Biotechnology (NY)* **6**(5), 419-429.

Yamauchi, M.M., Miya, M.U. & Nishida, M. (2004b) Use of a PCR-based approach for sequencing whole mitochondrial genomes of insects: two examples (cockroach and dragonfly) based on the method developed for decapod crustaceans. *Insect Molecular Biology* **13**(4), 435-442.

DECLARATION

I hereby declare that this thesis has not already been accepted in substance for any degree and is not being currently submitted in candidature for any other degree. It is the result of my own independent investigation and all authorities and sources that have been consulted are acknowledged in the bibliography

----

HOANG-DUNG TRAN

Date: August 01, 2009

CURRICULUM VITAE

Name: Hoang-Dung Tran

Date of birth: 31 March 1975

Marital status: Married (no children)

Nationality: Vietnam

Heimatadresse: 017A My-Thuan Apartment,
An-Duong-Vuong Street, F16 Q8, Hochiminh City, Vietnam

E-mail: tranhoangdung1975@yahoo.com

Phone: ++0084862606218

Hotline: ++00841222999537

High- Schools Studies

1981-1983: Khanh-Hoi A school, Hochiminh city, Vietnam

1983-1895: Tang-Bat-Ho B school, Hochiminh city, Vietnam

1985-1990: Tang-Bat-Ho A school, Hochiminh city, Vietnam

1990-1993: Le-Qui-Don high school, Hochiminh city, Vietnam

University Study

10.1993-09.1997: Bachelor of Science (Microbiology and Biochemistry), University of Natural Sciences, Vienam National University at Hochiminh City, Vietnam.

Promotions Studies

10.1997-09.2001: Master student of course of animal physiology at University of Natural Sciences, Vienam National University at Hochiminh City, Vietnam (Not finished).

10.2001-09.2002: International Training Course (ITC) Certificate, Biological Research Center, Hungarian Academy of Sciences, Szeged, Hungary.

10.2002- 09.2005: Fellowship from "*International Graduate School in Genetics and Functional Genomics*", University of Cologne, Germany.

05.2003-03.2005: Predoctoral thesis (equivalent to Diploma/Master Degree), *International Graduate School in Genetics and Functional Genomics,* Lab of Prof. Dr. Michael Melkonian, Institute of Botany, University of Cologne, Germany.

04.2003- 06.2006: Ph.D student, *International Graduate School in Genetics and Functional Genomics*, Lab of Prof. Dr. Michael Melkonian, Institute of Botany, University of Cologne, Germany.

Supervisor: Prof. Dr. Michael Melkonian Botanisches Institut, Lehrstuhl I Mathematisch-Naturwissenschaftliche Fakultat, Universität zu Köln, Gyrhofstr. 15 D-50931 Köln, Deutschland.

LEBENSLAUF

Name: Hoang-Dung Tran

Geburtsdatum: 31 March 1975

Familienstand: Married (no children)

Staatsangehörigkeit: Vietnam

Heimatadresse: 017A My-Thuan Apartment,
An-Duong-Vuong Street, F16 Q8, Hochiminh City, Vietnam

E-mail: tranhoangdung1975@yahoo.com

Phone: ++0084862606218

Hotline: ++00841222999537

Schulausbildung

1981-1983: Khanh-Hoi A school, Hochiminh city, Vietnam

1983-1895: Tang-Bat-Ho B school, Hochiminh city, Vietnam

1985-1990: Tang-Bat-Ho A school, Hochiminh city, Vietnam

1990-1993: Le-Qui-Don high school, Hochiminh city, Vietnam

Universität Studium

10.1993-09.1997: Bachelor of Science (Microbiology and Biochemistry), University of Natural Sciences, Vienam National University at Hochiminh City, Vietnam.

Promotions Studium

10.1997-09.2001: Master student of animal physiology course at University of Natural Sciences, Vienam National University at Hochiminh City, Vietnam (Not finished).

10.2001- 9.2002: International Training Course (ITC) Certificate, Biological Research Center, Hungarian Academy of Sciences, Szeged, Hungary.

10.2002- 09.2005: Stipendiat bei der "*International Graduate School in Genetics and Functional Genomics*", Universität zu Köln.

05.2003-03.2005: Predoctoral thesis (equivalent to Diploma/Master Degree), *International Graduate School in Genetics and Functional Genomics,* Prof. Dr. Michael Melkonian, Botanisches Institut, Lehrstuhl I, Mathematisch-Naturwissenschaftliche Fakultat, Universität zu Köln, Deutschland.

04.2005- 06.2006: Ph.D student, *International Graduate School in Genetics and Functional Genomics*, Prof. Dr. Michael Melkonian, Botanisches Institut, Lehrstuhl I, Mathematisch-Naturwissenschaftliche Fakultat, Universität zu Köln, Deutschland.

Betreurer: Prof. Dr. Michael Melkonian Botanisches Institut, Lehrstuhl I Mathematisch-Naturwissenschaftliche Fakultat, Universität zu Köln, Gyrhofstr. 15 D-50931 Köln, Deutschland.