Plastid genome structure and evolution: operons, novel genes and inteins

by

Shenglong Wang

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia, Canada
May, 1996

ISBN 0-612-16026-2

Name [redacted]

Dissertation Abstracts International and *Masters Abstracts International* are arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation or thesis. Enter the corresponding four-digit code in the spaces provided.

Biochemistry
SUBJECT TERM

0 4 8 7
SUBJECT CODE

UMI

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture	0729
Art History	0377
Cinema	0900
Dance	0378
Fine Arts	0357
Information Science	0723
Journalism	0391
Library Science	0399
Mass Communications	0708
Music	0413
Speech Communication	0459
Theater	0465

EDUCATION

General	0515
Administration	0514
Adult and Continuing	0516
Agricultural	0517
Art	0273
Bilingual and Multicultural	0282
Business	0688
Community College	0275
Curriculum and Instruction	0727
Early Childhood	0518
Elementary	0524
Finance	0277
Guidance and Counseling	0519
Health	0680
Higher	0745
History of	0520
Home Economics	0278
Industrial	0521
Language and Literature	0279
Mathematics	0280
Music	0522
Philosophy of	0998
Physical	0523
Psychology	0525
Reading	0535
Religious	0527
Sciences	0714
Secondary	0533
Social Sciences	0534
Sociology of	0340
Special	0529
Teacher Training	0530
Technology	0710
Tests and Measurements	0288
Vocational	0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language	
General	0679
Ancient	0289
Linguistics	0290
Modern	0291
Literature	
General	0401
Classical	0294
Comparative	0295
Medieval	0297
Modern	0298
African	0316
American	0591
Asian	0305
Canadian (English)	0352
Canadian (French)	0355
English	0593
Germanic	0311
Latin American	0312
Middle Eastern	0315
Romance	0313
Slavic and East European	0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy	0422
Religion	
General	0318
Biblical Studies	0321
Clergy	0319
History of	0320
Philosophy of	0322
Theology	0469

SOCIAL SCIENCES

American Studies	0323
Anthropology	
Archaeology	0324
Cultural	0326
Physical	0327
Business Administration	
General	0310
Accounting	0272
Banking	0770
Management	0454
Marketing	0338
Canadian Studies	0385
Economics	
General	0501
Agricultural	0503
Commerce-Business	0505
Finance	0508
History	0509
Labor	0510
Theory	0511
Folklore	0358
Geography	0366
Gerontology	0351
History	
General	0578
Ancient	0579
Medieval	0581
Modern	0582
Black	0328
African	0331
Asia, Australia and Oceania	0332
Canadian	0334
European	0335
Latin American	0336
Middle Eastern	0333
United States	0337
History of Science	0585
Law	0398
Political Science	
General	0615
International Law and Relations	0616
Public Administration	0617
Recreation	0814
Social Work	0452
Sociology	
General	0626
Criminology and Penology	0627
Demography	0938
Ethnic and Racial Studies	0631
Individual and Family Studies	0628
Industrial and Labor Relations	0629
Public and Social Welfare	0630
Social Structure and Development	0700
Theory and Methods	0344
Transportation	0709
Urban and Regional Planning	0999
Women's Studies	0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture	
General	0473
Agronomy	0285
Animal Culture and Nutrition	0475
Animal Pathology	0476
Food Science and Technology	0359
Forestry and Wildlife	0478
Plant Culture	0479
Plant Pathology	0480
Plant Physiology	0817
Range Management	0777
Wood Technology	0746
Biology	
General	0306
Anatomy	0287
Biostatistics	0308
Botany	0309
Cell	0379
Ecology	0329
Entomology	0353
Genetics	0369
Limnology	0793
Microbiology	0410
Molecular	0307
Neuroscience	0317
Oceanography	0416
Physiology	0433
Radiation	0821
Veterinary Science	0778
Zoology	0472
Biophysics	
General	0786
Medical	0760

EARTH SCIENCES

Biogeochemistry	0425
Geochemistry	0996
Geodesy	0370
Geology	0372
Geophysics	0373
Hydrology	0388
Mineralogy	0411
Paleobotany	0345
Paleoecology	0426
Paleontology	0418
Paleozoology	0985
Palynology	0427
Physical Geography	0368
Physical Oceanography	0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences	0768
Health Sciences	
General	0566
Audiology	0300
Chemotherapy	0992
Dentistry	0567
Education	0350
Hospital Management	0769
Human Development	0758
Immunology	0982
Medicine and Surgery	0564
Mental Health	0347
Nursing	0569
Nutrition	0570
Obstetrics and Gynecology	0380
Occupational Health and Therapy	0354
Ophthalmology	0381
Pathology	0571
Pharmacology	0419
Pharmacy	0572
Physical Therapy	0382
Public Health	0573
Radiology	0574
Recreation	0575
Speech Pathology	0460
Toxicology	0383
Home Economics	0386

PHYSICAL SCIENCES

Pure Sciences

Chemistry	
General	0485
Agricultural	0749
Analytical	0486
Biochemistry	0487
Inorganic	0488
Nuclear	0738
Organic	0490
Pharmaceutical	0491
Physical	0494
Polymer	0495
Radiation	0754
Mathematics	0405
Physics	
General	0605
Acoustics	0986
Astronomy and Astrophysics	0606
Atmospheric Science	0608
Atomic	0748
Electronics and Electricity	0607
Elementary Particles and High Energy	0798
Fluid and Plasma	0759
Molecular	0609
Nuclear	0610
Optics	0752
Radiation	0756
Solid State	0611
Statistics	0463

Applied Sciences

Applied Mechanics	0346
Computer Science	0984
Engineering	
General	0537
Aerospace	0538
Agricultural	0539
Automotive	0540
Biomedical	0541
Chemical	0542
Civil	0543
Electronics and Electrical	0544
Heat and Thermodynamics	0348
Hydraulic	0545
Industrial	0546
Marine	0547
Materials Science	0794
Mechanical	0548
Metallurgy	0743
Mining	0551
Nuclear	0552
Packaging	0549
Petroleum	0765
Sanitary and Municipal	0554
System Science	0790
Geotechnology	0428
Operations Research	0796
Plastics Technology	0795
Textile Technology	0994

PSYCHOLOGY

General	0621
Behavioral	0384
Clinical	0622
Developmental	0620
Experimental	0623
Industrial	0624
Personality	0625
Physiological	0989
Psychobiology	0349
Psychometrics	0632
Social	0451

For my parents, Aiying Zhou and Shougen Wang, and my wife, Ying Zhang.

Table of contents

Page

Table of contents v

List of Figures x

List of Tables xii

Abstract xiii

List of Abbreviations xiv

Acknowledgments xvi

Chapter I. Background 1

A. Endosymbiotic Origins of Cell Organelles 2

(i). Molecular phylogeny based on rRNA sequence data 4

(ii). Protein sequence data in support of the endosymbiont hypothesis 5

(iii). Plastid gene clusters 6

(iv). Secondary endosymbiosis in plastid evolution 6

(v). Origin of plastids: monophyletic or polyphyletic 7

(vi). Gen. transfer in plastid evolution 9

(vii). Data from mitochondria supporting the endosymbiont hypothesis 11

B. Ribosomal Protein Operons 14

(i). The *S10* and *spc* ribosomal protein operons in *E. coli* 15

(ii). The *S10* and *spc* ribosomal protein operons in plant chloroplast genome 17

(iii). The *S10* and *spc* ribosomal protein operons in cyanelle 19

(iv). The *S10* and *spc* ribosomal protein operons in *Guillardia theta* chloroplast 19

C. Group I Intron-encoded DNA Endonucleases 20

(i). Group I intron 20

(ii). Group I intron-encoded DNA endonucleases 22 (iii). DNA endonucleases are themselves mobile genetic elements 24

D. Protein Splicing 25

(i). Protein splicing in yeast 26

(ii). Protein splicing in archaebacteria 27

(iii). Protein splicing in eubacteria 28

(iv). Mechanism of protein splicing 29

(v). Inteins are DNA endonucleases 33

Chapter II. Materials and Methods 35

Materials 35

Methods 36

(i). Methods involved in DNA manipulation 36

a. Preparation of DNA 36

b. DNA digestion, fragment purification and ligation to vectors 37

c. Transformation of plasmid DNA into *E. coli* cells 38

(ii). DNA sequencing 39

a. Systematic deletion 39

b. Preparation of ssDNA for sequencing 40

c. Preparation of dsDNA for sequencing 41

d. DNA sequencing by the dideoxy-chain termination method 41

(iii). Hybridization 42

a. Northern blot 42

b. Southern blot 43

c. Probe labeling 43

d. Hybridization 44

(iv). Polymerase chain reaction (PCR) 44

a. Isolation of total RNA from *E. coli* 44

b. cDNA synthesis 45

c. PCR with Vent DNA polymerase 46

d. PCR with Taq and Vent DNA polymerases 46

(v). SDS-polyacrylamide gels for proteins 47

a. Preparative gel (SDS-PAGE) 47

b. Analytical gel (SDS-PAGE) 47

(vi). Isolation of 30S ribosome subunit from *Chlamydomonas reinhardtii* chloroplast 48

(vii). Protein purification 49

a. Protein expression 49

b. Protein purification by affinity column chromatography 49

c. Protein purification by electroelution 50

d. Factor Xa digestion of purified fusion protein 50

e. Transblot of protein from SDS polyacrylamide gels to NitroPlus or PVDF membrane 51

(viii). Antibodies and Western blot hybridization 52

a. Raising antibodies 52

b. Antibody purification 52

c. Western hybridization 53

Chapter III. Ribosomal Protein Operons in Plastid Genome of *Guillardia theta* 54

INTRODUCTION 54

RESULTS 58

A. DNA sequence determination and identification of ribosomal protein genes 58

B. Ribosomal protein gene organization, operon structure and evolution 71

DISCUSSION 98

Chapter IV. The Discovery of Three Novel Genes, the *acpA*, *hlpA*, and *dnaK* genes in the plastid genome of *Guillardia theta* 103

INTRODUCTION 103

RESULTS 105

A. Cloning and DNA sequence determination of a 2.8 kbp DNA fragment isolated from the *Guillardia theta* plastid genome 105

B. Identification of the acyl carrier protein gene (*acpA*) 105

C. Identification of the histone-like protein gene (*hlpA*) 112

D. Identification of the Hsp70-like heat shock protein gene (*dnaK*) 118

DISCUSSION 124

Chapter V. Characterization of Three Chloroplast Genes of Unusual Structure in *Chlamydomonas* and the Discovery of Inteins and Protein Splicing 128

INTRODUCTION 128

RESULTS 137

A. Identification of IS2 of *Chlamydomonas eugametos* ClpP as a degenerate intein that can be restored to protein splicing by a single amino acid substitution 137

a. Identification of intein-like features in the IS2 sequence 137

b. Characterization of the unmodified (wildtype) IS2 sequence 140

c. Site-directed mutagenesis of the IS2 sequence 141

d. Characterization of the modified IS2 (IS2-m) for its protein splicing activity 144

e. Purification of the 66 kDa protein as a putative spliced protein 150

f. Identification of the 66 kDa protein as spliced protein by micro-protein sequencing 155

B. Characterization of the IS1-containing ClpP protein of *Chlamydomonas reinhardtii* 161

C. Characterization of the *rps3* gene products of *Chlamydomonas reinhardtii* 165

a. Production of domain-specific antibodies 165

b. Purification and specificity test of the domain-specific antibodies 168

c. Western blot analysis of *rps3* gene products in *Chlamydomonas reinhardtii* 169

DISCUSSION 178

IS2 as a restorable degenerate intein 178

The IS1-containing ClpP as a fusion protein 185

The *rps3* gene in *Chlamydomonas reinhardtii* 188

VI. References 193

List of Figures

Page

Figure 3-1. Sc..ematic illustration of secondary endosymbiosis of *Guillardia theta* 56

Figure 3-2. Subcloning of DNA fragment from *Guillardia theta* plastid genome 60

Figure 3-3. The complete DNA sequence of an 8.3 kbp DNA cloned from *Guillardia theta* plastid genome and deduced protein sequences 63

Figure 3-4. Alignment of the deduced ribosomal protein sequences with their homologs in other organisms 73

Figure 3-5. Comparison of ribosomal protein gene organization in *Guillardia theta* with that in other organisms 90

Figure 3-6. *Guillardia theta* plastid ribosomal protein gene-specific probes 93

Figure 3-7. Northern hybridization of *Guillardia theta* total RNA with plastid ribosomal protein gene-specific probes 95

Figure 3-8. Rearrangement of ribosomal protein gene clusters in the *Guilardia theta* plastid genome relative to corresponding operons in *E. coli* 97

Figure 4-1. Gene map of a 2.8 kbp DNA fragment from *Guillardia theta* plastid genome that encodes *acpA*, *hlpA* and *dnaK* genes 107

Figure 4-2. The complete DNA sequence of the 2.8 kbp fragment and the deduced protein sequences 109

Figure 4-3. Comparison of deduced AcpA protein sequence of *Guillardia theta* with known acyl carrier protein sequences 114

Figure 4-4. Comparison of deduced HlpA protein sequence of *Guillardia theta* with bacterial histone-like protein sequences 116

Figure 4-5. Comparison of deduced DnaK protein sequence of *Guillardia theta* with heat shock protein Hsp70 family proteins 119

Figure 5-1. Schematic comparison of protein sequences of *Chlamydomonas reinhardtii* and *Chlamydomonas eugametos* ClpP with that of other organisms 132

Figure 5-2. Schematic comparison of protein sequence of *C. reinhardtii* Rps3 with that of *E. coli* 135

Figure 5-3. The alignment of *Chlamydomonas eugametos* ClpP IS2 sequence with those of inteins and intein-like sequences 139

Figure 5-4. Site mutation and gene disruption of *Chlamydomonas eugametos* IS2 143

Figure 5-5. Clones of wildtype, mutant, and disrupted *Chlamydomonas eugametos* IS2 in pMAL vector 146

Figure 5-6. The expression of wildtype, mutant, and disrupted *Chlamydomonas eugametos* IS2 in *E. coli* 149

Figure 5-7. Reverse transcription of *Chlamydomonas eugametos* IS2 and PCR amplification of cDNA 152

Figure 5-8. Purification of ClpP-related proteins from *E. coli* by amylose affinity column chromatography 154

Figure 5-9. Digestion of the putative spliced protein with protease Factor Xa 157

Figure 5-10. FPLC profile of putative spliced protein digested by trypsin 160

Figure 5-11. Subcloning of *Chlamydomonas reinhardtii clpP* gene 164

Figure 5-12. Subcloning of *Chlamydomonas reinhardtii rps3* gene 167

Figure 5-13. Subclones of *Chlamydomonas reinhardtii rps3* gene used in antibody purification and specificity testing 171

Figure 5-14. The specificity of domain-specific antibodies of *Chlamydomonas reinhardtii* Rps3——Western blot analysis 173

Figure 5-15. The expression of *rps3* gene in *Chlamydomonas reinhardtii* chloroplast——Western blot analyses 177

List of Tables

Page

Table 3-1. Intergenic regions in the *Guillardia theta S10/spc* operon-like gene cluster and the location of putative ribosomal binding sites 86

Table 3-2. The percentage of identity of ribosomal proteins between plastid genome of *Guillardia theta* and that of other organisms 87

Table 4-1. The percentage of identity between *Guillardia theta* plastid HlpA protein and eubacterial histone-like proteins 117

Table 4-2. The percentage of identity between *Guillardia theta* plastid DnaK protein and heat shock proteins of Hsp70 family 123

Abstract

Plastid genome structure and evolution were studied by analyzing a cryptomonad plastid believed to have originated through secondary endosymbiosis. Sequence determination of a 14-kbp DNA fragment from the*Guillardia theta* plastid genome revealed 22 genes. The arrangement of these genes is *acpA-hlpA-dnaK-rpl3-rpl4-rpl23-rpl2-rps19-rpl22-rps3-rpl16-rpl29-rps17-rpl14-rpl24-rpl5-rps8-rpl6-rpl18-rps5-secY-rpl36*. The *acpA*, *hlpA* and *dnaK* genes encode, respectively, an acyl carrier protein, a histone-like protein, and an Hsp70 family protein. All three genes were found for the first time in a cell organelle genome, and each gene represented a functional class of plastid genes not previously described. The downstream genes are ribosomal protein genes (except *secY*), and their organization resembles the two largest and adjacent ribosomal protein operons (*S10* and *spc*) of *E. coli*. An RNA transcript of 10,000 nucleotide long was detected for these ribosomal protein genes, further suggesting that they are expressed as an operon. When compared to corresponding operons (or gene clusters) in plastid genomes of other organisms, this *Guillardia theta* plastid ribosomal protein operon has retained the largest number of genes. Relative to corresponding ribosomal protein operons (*str-S10-spc-alpha*) in *E. coli*, the *Guillardia theta* plastid operons appear to have undergone rearrangement, expansion, and fusion in forming a much larger ribosomal protein operon.

The expression and protein products of two structurally unusual *Chlamydomonas* chloroplast genes were investigated for possible presence of inteins and protein splicing. They are the *rps3* gene encoding a ribosomal protein (Rps3) and the *clpP* gene encoding a protease protein (ClpP), with both genes containing large translated insertion sequences. For the ClpP protein, it was demonstrated that one of its insertion sequence, IS2, is a degenerate intein that can be restored to protein splicing by a single amino acid substitution. The other large insertion sequence in the ClpP protein, IS1, was not excised from the precursor protein either *in vitro* or *in vivo* under conditions studied, indicating that it is not an intein. Similarly, an insertion sequence in the Rps3 protein does not appear to be an intein. Instead, the Rps3 precursor protein appears to be cleaved into smaller protein products, all of which are assembled into the ribosomal small subunit. The IS2 sequence of the ClpP protein thus represents the first, and so far the only, intein found in an organelle genome, while the ClpP protein represents the fourth functionally distinctive protein found to contain an intein.

List of Abbreviations

A, absorbance

A, adenosine

aa, amino acid

C, cytidine

C, degree Celsius

dNTP, 2'-deoxynucleoside-5'-triphosphate

ddNTP, 2',3'-dideoxynucleoside-5'-triphosphate

DMSO, dimethylsulfoxide

DNA, deoxyribonucleic acid

dsDNA, double-stranded DNA

DTT, dithiothreitol

EDTA, ethylenediaminetetraacetate

EtBr, ethidium bromide

g, gravity

G, guanosine

h, hour

IPTG, isopropyl-ß-D-thiogalactopyranoside

kb, kilobase

kbp, kilobase pair

kDa, kiloDalton

LSU, large subunit

µg, microgram

µl, microliter

µM, micromolar

min, minute

ml, milliliter

mM, millimolar

M, molar

ng, nanogram

nt, nucleotide

PCR, polymerase chain reaction

PEG, polyethylene glycol

RNA, ribonucleic acid

rpm, revolutions per minute

RT, reverse transcription

SDS, sodium dodecyl sulfate

SDS-PAGE, SDS polyacrylamide gel electrophoresis

sec, second

SSU, small subunit

ssDNA, single stranded DNA

T, thymidine

TEMED, *N,N,N'N'*-tetramethylethylenediamine

Tris, tris(hydroxymethyl)-aminomethane

UV, ultraviolet

Acknowledgments

I take this opportunity to formally thank those people who have helped me, in one way or another, to meet the challenge of completing the research described in this thesis and in obtaining this degree.

I thank my supervisor, Dr. Paul X-Q. Liu for allowing me the opportunity to do research in his laboratory. I am particularly appreciative of the freedom given me to work on my own time schedule and of the supervisic.. .ıd patience that Dr. Liu has displayed during this time period, without his supportive attitude and encouragement, no single one of my ideas would have been realized. Gratitude is also extended to my supervisory committee members: Dr. W. F. Doolittle, Dr. R. A. Singer and Dr. M. W. Gray for their guidance and input as my research progressed.

I thank the external examiner, Dr. L. Bonen and the other members of the examining committee, Dr. R. Lee, Dr. W. F. Doolittle and Dr. M. W. Gray for their time spent in reviewing this thesis. From that I learnt a lot about how to write in English in addition to other issues.

I would like to thank Dr. S. E. Douglas for her kindness and friendliness. Collaboration with her in working with *Guillardia theta* was very pleasant and fruitful. I greatly appreciate for her time spent in reviewing my thesis and for her valuable comments and suggestions on various issues.

I would also like to thank the former and present members of Dr. Gray's and Dr. Lee's laboratories, with special reference to Dr. Murray Schare and Dr. David Spencer. I have been fortunate to have weekly joint-laboratory meetings with them, their comments and suggestions are greatly appreciated.

It is my great pleasure to extend my special "thank you" to my wife Ying Zhang, her unselfish sacrifice and constant support made it possible for me to concentrate on my research. She has given me and is continuously giving me the inspiration and motivation that is so important when a challenge of this magnitude is undertaken.

Chapter I. Background

Eukaryotic cells are fundamentally different from prokaryotic cells, not only because their genetic materials are bounded by membranes and they use different strategies to assert cellular functions, but also because they contain other functionally specialized subcellular structures (organelles). Prominent among organelles are plastids and mitochondria. Most (but not all) eukaryotes possess mitochondria which provide the cell with aerobic respiration through oxidative phosphorylation, resulting in the generation of ATP. Plastids are unique to photosynthetic eukaryotes, such as land plants and algae, which collect energy directly from the sun through photosynthesis. Plastids and mitochondria are membrane-bounded compartments within the cytoplasm of eukaryotic cells. Unlike other subcellular compartments (such as lysosome), plastids and mitochondria have their own genomes in the form of relatively small molecules of typically circular double-stranded DNA, from which proteins of distinct and specialized functions are produced. Ever since they were discovered over a hundred years ago, the question of how plastids and mitochondria arose has been a subject of continuing interest and debate in evolutionary molecular biology.

In this chapter, I will first discuss the endosymbiotic origins of cell organelles and the evidence supporting the endosymbiont hypothesis. Then I will discuss the origin of plastids in terms of whether a single endosymbiotic cyanobacterium was the ancestor of all plastids, or different photosynthetic endosymbionts were the ancestors of plastids in different groups of organisms (monophyletic versus polyphyletic). Following this I will briefly discuss secondary endosymbiosis, in which the endosymbiont was a free-living photosynthetic eukaryote rather than a prokaryote. In later sections, I will discuss the gene structure and organization in plastid genomes, particularly focusing on ribosomal protein operon-like gene clusters. And finally, I will discuss specific aspects of chloroplast gene expression that have particular relevance to other areas of the research project presented in this thesis, including DNA endonuclease-mediated intron homing and protein splicing.

A. Endosymbiotic origins of cell organelles

Eukaryotic cells are internally more complicated than prokaryotic cells. They contain a nucleus in which the genetic material (DNA) is to some extent separated from the cytosol content. They also contain many other subcellular compartments, most prominent among them being the organelles mitochondria and plastids. Organelles are unique in structure and specialized in function. Mitochondria oxidize energy-rich substrates through oxidative phosphorylation to produce ATP, and plastids capture light energy from the sun through photosynthesis. Organelles are unique also because they contain their own genomes, which are typically circular double-stranded DNA molecules. Mitochondrial genomes range in size from 14 kbp (some animals) to as much as 2400 kbp (some land plants), while plastid genome sizes are in a much narrower range, usually 100 to 200 kbp. Organelle genome sizes are relatively small, much smaller than a simple free-living eubacterium (such as *E. coli*). Even though organelles are specialized in certain functions (aerobic respiration for mitochondria and photosynthesis for plastids), the number of genes encoded in their genomes is far from enough for their specific functions, and many of the proteins required for these functions are encoded in the nucleus and transported into organelles subsequently. For these reasons, the origin of organelles and their evolution have been the subject of continuing debate ever since they were discovered.

There have been two major theories that seek to explain the origin of organelles in eukaryotic cells. One of them is the autogenous origin theory, which proposes that these organelles arose from within the cell through a process of intracellular compartmentalization and functional specialization. According to this theory (Raff and Mahler, 1972), the protoeukaryote (which gave rise to today's eukaryotes with mitochondria inside) was an advanced aerobic cell with large size. The increased respiratory membrane invaginated to form the organelle, followed by implantation of a circular DNA containing genes required for the functions of the organelle. The generation of the implanted DNA is similar to the generation of multiple nucleoli (small circular DNAmolecules specialized for the transcription of rRNA) during amphibian oogenesis.

The endosymbiont hypothesis assumes that the organelle was once a free-living prokaryote (Gray and Doolittle, 1982; Gray, 1992). When engulfed by a protoeukaryote, it entered into a symbiotic relationship with the host by providing it with aerobic respiration or photosynthesis. Soon after the establishment of symbiosis, a massive loss of the symbiont's genetic materials occurred through gene loss or gene transfer to the host's nuclear genome, which accounts for the small genomes of organelles relative to their ancestral eubacterial genomes. The organelle becomes more and more dependent on the nuclear genome of the host cell for its function as more and more organellar components had to be imported from the host's cytosol.

As both organelles (mitochondria and plastids) and the host have their own genomes, these two alternative theories can be tested by molecular biological analysis. In distinguishing the two opposing theories, three forms of evidence can be taken as the proof for the endosymbiont hypothesis (Gray and Doolittle, 1982): (i), if the evolutionary histories of nuclear genomes and one of the organellar genomes were known with certainty, and the two could be shown to derive from genomic lineages which were phylogenetically distinct before the formation of the eukaryotic cell; (ii), if the nuclear genome, although lacking modern free-living relatives, can be shown to have descended from a lineage other than that from which organellar genomes descended; and (iii), it could be shown that organelles of different major groups derived from different lineages of prokaryotes. From the time the endosymbiont hypothesis was proposed, more and more evidence has been obtained in support of this theory. As the evidence from mitochondria largely parallels that from plastids, I will focus the discussion mainly on the evidence from plastids.

(i). Molecular phylogeny based on rRNA sequence data

Ribosomal RNA sequence ccmparisons have played an important role in deciphering the pathways of organelle evolution. The features of SSU and LSU rRNAs are very valuable in building phylogenetic trees. Such features include antiquity (descent of all extant rRNA homologs from a common ancestor), ubiquity (the presence of homologous rRNA genes in all genome types—archaebacteria, eubacteria, eukaryotes, mitochondria, and plastids), functional equivalency and constancy, the presence of both slowly evolving and more rapidly evolving segments that facilitate the determination of relationships over a broad range of evolutionary distance, the ability of slow-evolving segments to form a highly conserved core of secondary structure that facilitates accurate alignment of primary sequences, the availability of a large number of nucleotide positions which minimizes statistical fluctuations, and the relative ease of complete or nearly complete sequence determinations. Phylogenetic trees based on these kinds of comparisons divide primary life into three monophyletic lineages. They are archaebacteria, eubacteria, and eukaryotes. Studies of the SSU rRNA gene of wheat mitochondria (Bonen et al., 1977) and the complete sequences of SSU rRNA gene from fungal and animal mitochondria all suggested that mitochondria were derived from eubacterial endosymbionts. Complete sequence of wheat mitochondrial SSU rRNA gene has shown that sequence similarity within the universal regions was significantly higher to eubacterial homolog than to nuclear homolog (Spencer et al., 1984). In global rRNA trees that define three monophyletic lineages of primary life (archaebacteria, eubacteria, and eukaryotes), plastids clearly fall within the eubacterial lineage and cluster specifically with cyanobacteria (Gray et al., 1984; Giovannoni et al., 1988; Cedergren et al., 1988; Turner et al., 1989). In rRNA trees, all types of plastids appear to be of cyanobacterial origin, and all are more closely related to one another than they are to a cyanobacterium such as *Anacystis nidulans* (Douglas and Turner, 1991).

(ii). Protein sequence data in support of the endosymbiont hypothesis

Phylogenetic trees based on protein sequence data also support a cyanobacterial origin of plastids; these data include cytochrome c gene sequence (Schwartz and Dayhoff, 1981), ATP synthase subunit 6 (Cozens et al., 1986), *tufA* (encoding translational elongation factor Tu) (Baldauf and Palmer, 1990), and RNA polymerase C1 subunit (*rpoC1*) (Palenik and Haselkorn, 1992). Protein trees based on *rbcS* and *rbcL* genes show surprising discrepancy with rRNA trees in that these genes appear specifically related to their homologs in ß-purple bacteria, not cyanobacteria. Similar observations have been made for the Rubisco proteins from different groups of plastids (chromophyte and rhodophyte algae versus green algae and land plants). Although it has been suggested that the discrepancy between rRNA and protein data supports a polyphyletic origin of plastids, a noncyanobacterial origin of rhodophyte plastids is incompatible with their strong resemblance to cyanobacteria in other aspects such as pigment composition and photosynthetic membrane structure. It seems more likely that rhodophyte and chromophyte plastids obtained the *rbcS-rbcL* operon from a β-purple bacterium through a process of horizontal gene transfer.

(iii). Plastid gene clusters

One aspect of plastid genome organization that can be useful in studying plastid origin and evolution is the clustering of genes or operons. A number of plastid gene operons, such as *atpBE, psbEFLJ, psbDC,* and *psaAB*, are conserved in all the plastids investig`ted so far, with the exception of chlorophytes (Boudreau et al., 1994). These gene clusters are present in cyanobacteria, reflecting the ancestral organization. Some gene clusters are conserved in all the plastids but differ from the organization of corresponding genes in cyanobacteria, indicating a common ancestor of all plastid types. Two such kind of gene clusters have been identified (Reith and Munholland, 1993), and several more are possible candidates.

Several gene clusters may indicate plastid-specific gene arrangements, but the absence of data from cyanobacteria and some plastids, and gene transfer to the nucleus, make the analysis of these gene clusters more complicated. One of these gene clusters is a large ribosomal protein operon that contains homologues of the *E. coli S10, spc, alpha* and *str* operons. This operon is present in *P. purpurea* and chromophytes, but it is organized differently in *C. paradoxa,* chlorophyte, and metaphyte. Further analyses of cyanobacterial as well as chromophyte and chlorophyte plastid genomes are required to clarify the evolution of these gene clusters in plastids.

(iv). Secondary endosymbiosis in plastid evolution

The role of secondary endosymbiosis in plastid evolution should be considered before assessing evidence regarding plastid origins. The evidence of secondary endosymbiosis has come from phylogenetic studies of 18S rRNA gene sequences of *Guillardia theta*. The nucleomorph-containing cryptomonad algae currently constitute the most convincing evidence of secondary acquisition of plastids from a eukaryotic algal endosymbiont. It has been shown that the nucleomorph, which is assumed to be the vestigial nucleus of such an endosymbiont, contains DNA (Ludwig and Gibbs, 1985). McFadden (1990) demonstrated more recently that eukaryotic nuclear-type rRNA genes are present in a nucleolus-like structure in the nucleomorph of the cryptomonad *Chroomonas caudata*. Eschbach et al. (1991) have shown that the nucleomorph of another cryptomonad, *Pyrenomonas salina*, contains three linear chromosomes, all of which hybridize with eukaryotic-type SSU and LSU rDNA probes, indicating that nucleomorph DNA may encode rRNA in cryptomonads. The complete sequences of two distinctive eukaryotic-type 18S rRNA genes (Nu and Nm for nuclear and nucleomorph) in cryptomonad algae *Guillardia theta* (Douglas et al., 1991) and the subsequent phylogenetic analysis of these sequences clearly demonstrated that the Nm sequence clustered specifically with the 18S rRNA sequences of rhodophytes, whereas the Nu sequence clustered with the assemblage containing chlorophytes and metaphytes, suggesting that the nucleomorph was derived from the nucleus of a rhodophyte endosymbiont. Recent *in situ* hybridization studies (McFadden et al., 1994a) further demonstrated that the Nm 18S rRNA is specifically associated with the nucleomorph. A similar *in situ* hybridization study has been conducted on the rRNA genes of a chlorarachniophyte (McFadden et al., 1994b), which is an amoeboid alga with plastids that are bounded by four membranes and containing a nucleomorph. The data obtained from cryptomonads and chlorarachniophytes strongly suggested that plastids in these groups were obtained through secondary endosymbiosis from eukaryotic endosymbionts. By extension, all other plastids bounded by more than two membranes would have been obtained the same way. This would include plastids of other chromophytes and euglenoids, as the plastids of this latter group are bounded by three membranes.

(v). Origin of plastids: monophyletic or polyphyletic

As discussed above, it is almost certain that plastids have originated through a series of separate endosymbioses in different branches of the eukaryotic tree (Gray, 1992), and in some cases secondary endosymbiosis was involved. From this perspective, plastids can be considered to have had multiple origins. However, a more fundamental issue is whether the plastids of all plants and algae can trace their existence to the same primary endosymbiotic event involving a single eubacteria-like ancestor (monophyletic origin), or whether plastids arose separately in different eukaryotic groups as a result of independent associations with eubacteria-like progenitors (polyphyletic origin). If all chromophytes, euglenoids, and chlorarachniophytes arose through secondary endosymbiosis involving rhodophytes or chlorophytes, the debate between monophyletic and polyphyletic origins of plastids is reduced to distinguishing between monophyletic and diphyletic options (chlorophytes versus rhodophytes) (Reith, 1995). The observation that most phylogenetic trees associate all plastids with cyanobacteria strongly supports a monophyletic origin of plastids. Other evidence in support of a monophyletic origin of plastids includes plastid operons, functional interchangeability of transit peptides and conserved GapA transit peptide and the first intron (which suggests that both rhodophyte and metaphyte plastids are derived from a single primary endosymbiotic event) (Reith, 1995). On the other hand, there is substantial evidence against a polyphyletic origin. Prochlorophytes are polyphyletic within the cyanobacteria, but phylogenetic trees based on rRNA gene and protein sequences show no specific affiliation with chlorophyte or metaphyte plastids. Similarly, phylogenetic analyses do not support the suggestion that chromophyte plastids were derived from *Heliobacterium chlorum*. A seven-amino-acid gap at the carboxyl terminus of PsbA was found in chlorophytes, metaphytes, some species of prochlorophytes, but not in rhodophytes, chromophytes and other prochlorophyte species. This characteristic has often been interpreted as supporting a polyphyletic origin of plastids, but phylogenetic trees of PsbA do not show a clear association of prochlorophytes with chlorophytes and metaphytes. Although the issue has not been definitively resolved yet, current evidence points toward a monophyletic origin of plastids.

(vi). Gene transfer in plastid evolution

Although all the molecular data suggest that organelles evolved from eubacterial endosymbionts, organellar genomes are substantially smaller than their ancestral eubacterial genomes. For instance, the largest chloroplast genome known so far is about 400 kbp, the smallest animal mitochondrial genome is only 14 kbp, whereas the smallest cyanobacterial genome is 3,100 kbp. The number of genes encoded in plastid genomes varies from above 100 to less than 300, not even enough for their specialized functions (i.e. gene expression and photosynthesis). Obviously, such small-sized genomes would not be able to sustain an ancestral free-living organism as the endosymbiont hypothesis would predict.

However, this genome size discrepancy can be well accommodated within the endosyn)iont hypothesis, which assumes that a massive loss and transfer of genetic infcrmation from the endosymbiont genome to the nucleus occurred in the course of organellar evolution. In fact, the majority of the genes specifying organellar structure and function are nuclear genes, and their products (mostly proteins) are imported back into the respective organelles where they function in combination with organellar gene products in executing the specialized functions of the organelles. Some of these genes may have already been present in the nuclear genome before the endosymbiosis and happened to duplicate genes in the endosymbiont genomes, so that the redundant endosymbiont genes could be lost. Many others, however, were almost certainly transferred from the endosymbiont genome to the host nuclear genome. Several pathways and mechanisms of such gene transfer have been postulated. In general, the process of gene transfer involves duplication of endosymbiont gene in the host genome (reverse transcribed and integrated into host genome), acquisition of function by the nuclear copy of the gene (acquisition of promoter and sometimes introns) (at this stage both the nuclear and organellar genes may have been active), selection of the nuclear gene over the organellar gene for its ultimate function (including finding a way back into the organelle), and eventually, loss of the now silent organellar gene (deletion). Evidence supporting this scenario came from the sequence analysis of nuclear genes encoding the cytosolic and plastid forms of glyceraldehyde 3-phosphate dehydrogenase (GAPDH). Both subunits of the chloroplast GAPDH are encoded in the nucleus, but their amino acid sequences are more closely related to those of thermophilic eubacteria than to that of cytosolic GAPDH present in the same cell (Shih et al., 1986; Martin and Cerff, 1986; Brinkmann et al., 1987; Liaud et al., 1990).

Because of the general conservation of plastid genome size and the observation that all mitochondrial and plastid genomes generally encode a similar set of genes, most of the gene transfer must have occurred relatively early in the evolution, i.e., soon after the endosymbiosis took place. However, the fact that some genes are encoded in plastid genomes in some species but in nuclear genomes in other species suggests that the gene transfer must be an ongoing process. Strong evidence supporting this scenario came from study of *tufA* gene (encoding a chloroplast protein synthesis elongation factor Tu) in land plants (Baldauf and Palmer, 1990). The *tufA* gene is encoded in the nucleus in the land plant *Arabidopsis thaliana* but phylogenetic trees of this gene clearly clustered it with the cyanobacterial-like plastid-encoded genes of green algae, suggesting that the plastid-to-nucleus transfer of this gene occurred after the separation of metaphytes from chlorophytes. Subsequently, a copy of that *tufA* gene was found in the plastid genome of one species in the charophyte lineage, although generally this lineage lacks a *tufA* gene in the plastid genome. Moreover, the plastid-encoded gene is so diverged that it is very unlikely that the gene product is still able to act as a functional EF-Tu.

Evidence of more recent gene transfer has also been described. The large subunit ribosomal protein gene *rpl22* has been transferred from the chloroplast genome to the nuclear genome within a common ancestor of flowering plants (Gantt et al., 1991). The Rubisco SSU gene was found in the plastid genome of rhodophytes and chromophytes but in the nuclear genome of chlorophytes and metaphytes (Palmer, 1985). Finally the *cox2* gene has been transferred from the mitochondrial genome to the nuclear genome sometime after the separation of monocots and dicots (Covello and Gray, 1992). Furthermore, the examples of recent gene transfer do not seem to be restricted to the chlorophyte/metaphyte lineage. In fungi, the functional *atp9* gene is encoded in mitochondrial genome of *Saccharomyces*, but in *Neurospora* and *Podospora*, this gene is in the nucleus, and the *atp* gene in *Neurospora* mitochondria is silent (van den Boogaart et al., 1982; Brown et al., 1985).

(vii). Data from mitochondria supporting the endosymbiont hypothesis

Since the endosymbiont hypothesis was proposed (Gray and Doolittle, 1982), a substantial amount of data has been obtained from plastid molecular biology to support the hypothesis of an endosymbiotic, cyanobacterial origin of plastids, and the alternative hypothesis (autogenous) of plastid origin no longer demands serious consideration. During the same period of time, emerging mitochondrial molecular biology has not provided comparably overwhelming support for the long-standing proposal of an endosymbiotic, eubacterial origin of mitochondria. Part of the reason is that the patterns of mitochondrial genome organization and gene expression are so diverged that in very few cases are these features obviously eubacterial, and molecular biology in mitochondria is in many of its specifics unlike anything seen in eubacteria, archaebacteria or the eukaryotic nucleus. More than that, mitochondria have a lot of novel, sometimes bizarre, features. These unusual features include deviation from the universal genetic code, unusual tRNA structures and codon recognition patterns, distinctive modes of gene organization and gene expression, promiscuous DNA, mosaic genes, RNA editing, scrambled rRNA genes, RNA import, and most recently, trans-splicing (Gray, 1992).

Mitochondrial genomes vary in size over a more than 150-fold range, from as little as 13.8 kbp in the nematode worm *Caenorhabditis elegans* to as much as 2400 kbp in muskmelon *Cucumis melo*. Despite the two orders of magnitude size difference between the smallest and the largest mitochondrial genomes, there is no indication of a corresponding difference in gene content. Throughout the range of mitochondria-containing eukaryotes, mtDNA has the same fundamental role: it specifies certain components of a mitochondrial protein-synthesizing system whose purpose is to translate a limited number of mitochondrially transcribed mRNAs that encode polypeptide components of the mitochondrial electron transport system. The respiratory chain genes are remarkably similar among structurally diverse mtDNAs, but the translation apparatus genes are more variable in distribution, with ribosomal protein genes completely or almost completely lacking in animal and fungal mtDNAs, but more frequent in the mitochondrial genomes of land plants and some protists. Nevertheless, analysis of homologous ribosomal protein genes showed that mitochondrial DNA and plastid DNA are about equally divergent from their *E. coli* homologs and from each other (Wahleithner and Wolstenholme, 1988), indicating that the mitochondrial versions are eubacterial in origin and the mitochondrial and plastid genes have had separate origins.

Other evidence supporting the endosymbiotic, eubacterial origin of mitochondria includes the observation that the translation system in mitochondria is more similar to the system in eubacteria than to those in archaebacteria and eukaryotic cytosol. Mitochondrial protein synthesis uses formylated Met-tRNA$^{\mathrm{Met}}$ as initiator and is inhibited by chloramphenical as in eubacteria; in contrast, in archaebacteria and eukaryotic cytosol, unformylated Met-tRNA$^{\mathrm{Met}}$ is used as an initiator, and protein synthesis is inhibited by cycloheximide. Features of the mitochondrial respiratory chain and oxidative phosphorylation resemble more closely to those found in the alpha-subdivision of the nonsulfur purple eubacteria than those of other aerobic bacteria (Woese et al., 1984). Moreover, phylogenetic trees based on *atp9* (nucleotide) sequence comparisons (Recipon et al., 1992) clearly show that all mitochondrial/nuclear *atp9* sequences cluster specifically with that of the alpha-purple bacterium *Rhodospirillum rubrum*.

Some of the most compelling evidence supporting eubacterial, endosymbiotic origin of mitochondria has come from studies of mitochondrial rRNA sequences, particularly those in land plants. Mitochondrial rRNA sequences from ciliates, fungi, and particularly animals have many unique primary and secondary structural features (Gutell et al., 1990) that makes it very difficult to assess their degree of similarity with one another and with their nonmitochondrial homologs (Gray, 1988). In striking contrast, plant mitochondrial SSU and LSU rRNAs show a remarkably high degree of primary sequence and secondary structural correspondence with their prokaryotic, and specifically eubacterial, counterparts. Moreover, wheat mitochondrial SSU rRNA has been shown to contain eubacteria-specific post-transcriptional modifications that have so far not been found in its mitochondrial homologs from other eukaryotes (Schnare and Gray, 1982; Gray, 1988). Plant mitochondrial ribosomes resemble prokaryotic 70S ribosomes rather than 80S eukaryotic ribosomes, although their rRNAs are bigger (eukaryotic sizes). The size differences are mainly due to the insertion of additional sequences into a few variable regions. The phylogenetic trees based on rRNA sequence comparisons not only have shown that plant mitochondria and eubacteria are specific evolutionary relatives, but have also made it possible to identify which particular eubacteria are the closest contemporary relatives of mitochondria. These turn out to be members of the alpha-subdivision of the purple bacteria.

B. Ribosomal protein operons

A fundamental difference between prokaryotes and eukaryotes in gene expression and organization is that in prokaryotes functionally related genes are organized into a single transcriptional unit known as an operon. For example in *E. coli*, the translation machinery known as the ribosome contains three rRNA molecules and 53 ribosomal proteins. As there are so many components devoted mainly to a single purpose, the protein translation for the synthesis of all the components has to be coordinately and stoichiometrically balanced. Organizing ribosomal protein genes into operons certainly contributed a lot in achieving this task, although the actual genetic organization of ribosomal protein operons in *E. coli* is complex, many of the ribosomal protein operons contain genes that are not ribosomal, and many of them are not typical operons in structure (two promoters or internal promoters).

In organelle evolution, the endosymbiont hypothesis proposes that cell organelles are derived from the eubacterial ancestors (mitochondria from alpha purple bacteria, plastids from cyanobacteria). According to this hypothesis, gene organization in the organellar genome should reflect their eubacterial ancestry. As we shall see in Chapter III, the organization of ribosomal protein genes in the plastid genomes does indeed resemble its eubacterial counterpart in operon structure. The operon-like gene regulation in organellar genomes may be important, especially during early stages of organellar evolution. After endosymbiosis, some genes in operons are transferred (copied) to the nucleus, and their organellar copies are deleted instead of left to accumulate mutations, which can partially account for the size shrinkage of organellar genomes. As a result, the flanking genes are brought together into close proximity, which is necessary if the operon-like regulation (translational coupling) is to be maintained. As more and more genes from the operon are transferred to the nucleus, including genes that are critical to the operon-like regulation, the structural bases (and thus the mechanisms) for regulating the genes as one unit (operon) are lost. Eventually, this may lead to the complete disappearance of the opeion in organelles.

A phylogenetic trees of a gene sequence has proven a useful tool in deciphering the aistory of organellar evolution, but it has an intrinsic risk. Precisely, a phylogenetic tree based on a particular gene sequence is a tree of that particular gene, not a tree of the genome encou.ng it (Gray, 1992). To what extent the two coincide depends on whether the gene in question has been subject to any sort of lateral transfer between genomes during its evolution, and whether the genome in question has acquired genes from other sources in the course of evolution. In this sense, study of organellar gene operons may be more valuable in dealing with these problems and may provide profound insight into evolutionary relationships.

(i). The *S10* and *spc* ribosomal protein operons in *E. coli*

Biosynthesis of ribosomes in *E. coli* is a substantial event such that under the most favorable laboratory growth conditions, nearly 40% of the total dried cellular mass of this bacterium is ribosomes (Nomura et al., 1984). With such a large amount of energy devoted to this event, it is essential that the synthesis of multiple ribosomal components be well balanced and precisely regulated. The equimolar synthesis of all the ribosomal components is achieved partly by organizing several component genes into one single transcriptional unit, known as an operon. All the genes within an operon are coordinately regulated and translationally coupled, thereby assuring the equimolar synthesis of all the components from one operon.

The *S10* and *spc* operons are the two largest operons in *E. coli*. The *S10* operon contains 11 ribosomal protein genes and *spc* operon contains 11 ribosomal protein genes and a *secY* gene that encodes a protein functioning in targeting and protein translocation across the plasma membrane. While their transcription is under the control of a typical *E. coli* promoters, these operons are further regulated at the post-transcriptional level by a mechanism known as feedback regulation, in which one of the proteins translated from an operon is used as a translational repressor to block the translation of its own mRNA. In the *S10* operon, the third gene product, Rpl4, can bind to the operon's translation initiation site to block the translation of all the genes encoded in this operon; while in the *spc* operon, the fifth gene product, Rps8, represses the translation of the third gene *rpl5* and downstream genes but not the first two genes, *rpl14* and *rpl24*, immediately following the promoter. Rpl4 also binds to the *S10* operon mRNA at a site preceding the translation initiation site of the first gene and causes transcription termination of the mRNA.

When a repressor ribosomal protein binds to its own mRNA, it does not always cause the degradation of the mRNA, and it does not even affect the transcription level of that mRNA. The regulation of the translation of genes downstream from that single target site is achieved by translational coupling, such that the translation of the genes downstream of an operon is dependent upon the translation of the first gene in that operon. In many operons demonstrating translational coupling, the independent initiation of translation from internal cistrons is effectively prevented, probably due to the lack of ribosome binding sites in between, or the translation initiation signal for the regulated downstream gene is normally sequestered in an RNA secondary structure and unavailable for recognition by ribosomes. On the other hand, the intercistronic distance between the two coupled cistrons is relatively small, so the translation reinitiation of the second cistron by the same ribosome is almost always certain to happen, thus guaranteeing the equimolar synthesis of both proteins.

It is worth mentioning that both repressor ribosomal proteins Rpl4 and Rps8 are known to bind to ribosomal RNAs strongly and specifically during early steps of ribosome assembly. In their role in feedback regulation, they recognize the same general structural features in their mRNAs that they recognize in rRNAs (Nomura et al., 1980). So the regulation of ribosomal protein synthesis is the result of a competition between similar binding sites on rRNA and ribosomal protein mRNA for repressor ribosomal protein. When rRNA is available, repressor ribosomal protein preferentially binds to rRNA; when it is not, the free repressor ribosomal protein binds to its own mRNA to prevent the overproduction of itself as well as other proteins encoded in the same operon.

(ii). The *S10* and *spc* ribosomal protein operons in plant chloroplast genome

Chloroplasts are plastids that perform photosynthesis within eukaryotes . They are autonomous replicons and have their own protein synthesis apparatus. Chloroplast ribosomes, as characterized by 2D-PAGE analysis in *Chlamydomonas, Euglena* and several land plants (Subramanian et al., 1991), are estimated to contain more than 60 ribosomal proteins, significantly more than the number of ribosomal proteins identified in *E. coli.* The chloroplast translation system is known to resemble the eubacterial translation system, but the genes are distributed between two genome compartments, namely the nucleus and chloroplast. In higher plant chloroplast genomes studied so far, approximately one-third of the ribosomal proteins are encoded in the chloroplast genome and the remaining two-thirds are encoded in the nucleus. Those genes encoded in nucleus are believed to have been transferred from the chloroplast genome to the nucleus after the endosymbiosis of a free-living photosynthetic organism, likely a cyanobacterium, which gave rise to the chloroplast we see today. The nuclear-encoded proteins are imported back to the chloroplast. Those genes still encoded in the chloroplast genome are organized into operons resembling those in eubacteria. Of the extra ribosomal proteins identified in the chloroplast ribosome that have no counterparts in *E. coli*, at least five of them have been characterized (Subramanian et al.. 1991, Wittmann, 1982, Gantt, 1988, Zhou and Mache, 1989, Johnson et al., 1990, Schmidt et al., 1993), and they all are encoded in the nucleus.

A total of 21 ribosomal protein genes have been identified in the completely sequenced chloroplast genomes of tobacco (Shinozaki et al., 1986), liverwort (Ohyama et al., 1986), rice (Hiratsuka et al., 1989), and maize (Weglöhner and Subramanian, 1993). These 21 ribosomal protein genes are all retained in all the four species, except that in the liverwort chloroplast genome *rps16* is absent but *rpl21* is present. Of the ribosomal proteins encoded by these 21 genes, 12 are from the small ribosomal subunit and 9 are from the large ribosomal subunit. It is interesting to note that the small ribosomal subunit, which makes the initial mRNA recognition and selects the correct position to start translation, is better represented. Almost all the 21 ribosomal proteins encoded in the chloroplast genome are homologues of *E. coli* ribosome early assembly proteins. In *E. coli*, these ribosomal protein genes all appear to be essential, as there is no *E. coli* mutant alive without any one of them except *rpl33*. In the chloroplast genome of maize (*Zea mays*), the largest ribosomal protein operon is L23-II (Weglöhner and Subramanian, 1993). It contains 11 ribosomal protein genes encoded in *E. coli* by three ribosomal protein operons (*S10, spc*, and *alpha*). It seems that the L23-II ribosomal protein operon is the fusion product of these three operons, but the other possibility (that the ancestor of the chloroplast encoded all the genes in one single operon but in *E. coli*, they are divided into three independent operons) can not be ruled out. It is also interesting to note that although chloroplast ribosomes contain a large number of acidic ribosomal proteins, only the basic or highly basic ribosomal proteins are encoded in the chloroplast genome (Subramanian et al., 1991). The significance of this preferential transfer is not clear.

(iii). The *S10* and *spc* ribosomal protein operons in cyanelle

The cyanelle is the photosynthetic organelle in cyanelle-containing photosynthetic organisms such as *Cyanophora paradoxa*. It is the equivalent of the chloroplast in higher plants, but the structural and biochemical characteristics are essentially cyanobacterial and close to the plastids of red algae. The cyanelle genome of *Cyanophora paradoxa* is studied the most relative to that of other cyanelle-containing organisms. Its genome size is 134 kbp, much smaller than the smallest genomes of known cyanobacteria, but similar to those of algal and higher plant plastids. The *S10* and *spc* ribosomal protein operons have been sequenced (Michalowski et al., 1990). The *S10/spc* region of the cyanelle genome contains 13 ribosomal protein genes, five more than in higher plant chloroplasts and four more than in *Euglena gracilis* plastids. Northern blots show that all the genes in this region are co-transcribed and then probably processed. The *rpS10* gene is the first gene in the *S10* operon in *E. coli*, and immediately upstream of the *rpS10* gene are sequences important for the post-transcriptional regulation of that operon. In cyanelles, the *rpS10* gene is located at the 3' end of the *str* operon, indicating that the regulation of the *S10/spc* operon may be different from that in *E. coli*.

(iv). The *S10* and *spc* ribosomal protein operons in *Guillardia theta* chloroplast

Guillardia theta is a unicellular organism with an unusual subcellular structure. While most photosynthetic eukaryotes acquired their photosynthetic apparatus through endosymbiosis of a prokaryote, the *Guillardia theta* apparently acquired its photosynthetic apparatus through endosymbiosis with another photosynthetic eukaryote (McFadden, 1990, Douglas et al., 1991). The plastid in cryptomonads is surrounded by four membranes (Gibbs, 1981), and between the inner and outer pairs of membranes there are ribosome-like particles and an apparently degenerated nucleus known as a nucleomorph (Greenwood, 1974). It is further demonstrated that rRNA species from nucleomorph and that from nucleus are both eukaryotic but are phylogenetically distinctive (Dougl·ᵉ et al., 1991). Little is known about the organization and arrangement of ribosomal protein genes in the plastid genome of *Guillardia theta*. Studies of the *str*-like ribosomal protein operon in this genome revealed that although gene organization of the *str*-like operon is similar to that of prokaryotes, the operon contains genes that are components of adjacent operons in their prokaryotic ancestor (Douglas, 1991). For part of this thesis, I sequenced a 14-kbp DNA fragment from the plastid genome of *Guillardia theta*, and 23 genes were idenıified. They include three non-ribosomal protein genes that encode, respectively, a Hsp70-like protein (DnaK), a histone-like protein (HlpA), and an acyl carrier protein (AcpA), and all of them were found in cell organellar genome for the first time (Wang and Liu, 1991). Downstream of these genes are 18 ribosomal protein genes and a *secY* gene which encodes a protein involved in protein translocation. These genes and their organization resemble prokaryotic *S10* and *spc* operons. When compared with other known plastid and cyanelle genomes, the *Guillardia theta* plastid genome has the most complete bacterium-like ribosomal protein operons. Northern blot analysis revealed a single messenger RNA of 10 kb that encodes all these genes, indicating that these genes are co-transcribed. This is in contrast to *E. coli* where the equivalent genes belong to two operons (*S10* and *spc*) that are normally transcribed into two separate polycistronic messenger RNAs.

C. Group I intron-encoded DNA endonucleases

(i). Group I intron

It was once thought that DNA carries all the information of a functional protein and RNA is an exact copy of information in one of the DNA strands. This underlying principle of the classical molecular biology was shattered forever with the discovery of RNA splicing. Now it has been known that many genes from virtually all the three major kingdoms have their coding sequences interrupted by stretches of noncoding DNA called intervening sequences or introns. Transcripts of such genes must undergo a cleavage-ligation process called RNA splicing to produce a mature, functional form of the mRNA, rRNA, or tRNA. In the decade since RNA splicing was first described, four major categories of splicing have been recognized based on their differentiated mechanisms and their unique RNA sequences and structures that contribute to the splicing reactions. They are group I, group II, nuclear mRNA and tRNA introns. Among them, the group I intron is particularly interesting. Besides its unique secondary and tertiary structures, the group I intron has at least two characteristic features very different from other types of introns, namely the mechanism of intron splicing and intron mobility. The splicing of the group I intron is initiated by the attack of the 3'-OH of a guanosine or one of its 5'-phosphorylated forms (GMP, GDP, or GTP) to the phosphorus atom of the 5' splicing site of the intron. The guanosine or one of its 5'-phosphorylated forms attackes to the intron through a 3',5'-phosphodiester bond and becomes the first nucleotide of the intron (Belfort, 1990; Cech, 1990). This is unique to the group I intron and mechanistically different from other introns. The mobility of group I introns is also unique. Unlike other transposable elements that involve non-homologous donor and recipient sites, group I intron mobility is site-specific in that it is restricted to exchanges between alleles of genes that contain or lack the intron (Dujon, 1989). This process, previously referred to as unidirectional intron conversion or site-specific intron transposition, is now termed "intron homing". A further distinction between group I intron mobility and most "conventional" transpositions is that the recipient DNA sequences that flank the inserted intron are not duplicated.

Site-specific recombination is not the only mechanism whereby introns can be inserted into compatible loci. Spliced introns could potentially insert themselves back into other RNA species by reversal of transesterification reactions used for splicing. Such recombined RNAs could than reintegrate into the genome, either by reverse transcription or by integrating directly into RNA or single stranded DNA at a replication fork.

Complete reversal of the splicing reaction has been demonstrated *in vitro* for the *Tetrahymena* group I intron (Woodson and Cech, 1989). During intron splicing, the internal guide sequences in the intron base-pair with the exon sequences at the splicing junctions, ensuring accurate splicing. Reversal of splicing may well use the same interactions that function in splicing. Maturases and other proteins that promote splicing could in principle also promote insertion of introns by reversal of splicing.

(ii). Group I intron-encoded DNA endonucleases

Among group I intron sequences published so far, most contain an ORF capable of encoding a protein. The group I intron-encoded protein can be expressed in three different ways: a, it may be translated from unspliced pre-mRNA as a fusion protein with the upstream exon; b, it may be translated from a separate mRNA species, possibly a processed form of the excised intron; and c, it may be expressed independently from its own promoter located within the intron. Of these group I intron-encoded proteins, some have been shown to be site-specific DNA endonucleases. The first such example was omega (Jacquier and Dujon, 1985), a 1.1-kb intron found in the large (21S) rRNA gene of mitochondrial DNA of some, but not all, strains of *Saccharomyces cerevisiae*. When a strain containing the intron is mated with a strain that does not contain it, the intron-less allele is nearly quantitatively converted into an intron-containing allele. Insight to the mechanism of the process was obtained by the finding that a transient double-stranded break appears at the intron homing site in intron-less DNA in zygotes derived from matings between the two cells (Zinn and Butow, 1985). Subsequent studies revealed that the expression of an open reading frame located within the intron is essential for both the DNA cleavage and the intron conversion. Detailed analysis of this gene conversion mechanism revealed that it is in many respects similar to the yeast HO endonuclease. In yeast the mating type switch is mediated by a site-specific endonuclease termed HO endonuclease, which is encoded by a gene unlinked to the mating type locus. In the gene conversion events, both endonucleases (HO endonuclease and group I intron-encoded endonuclease) make a double-stranded break at the target site, leaving a 4-nucleotide overhang with 3' termini, to promote gene conversion by a double-stranded break repair mechanism. These enzymes are different from type II restriction endonucleases in that their recognition sequences are longer (up to 18 nucleotide) and nonsymmetrical. Most of all, both contain the characteristic dodecapeptide LAGLI-DADG motif, which is believed to be the active site of the endonucleases. Significantly, this dodecapeptide motif is also found in most other group I intron-encoded ORFs.

The *td* and *sunY* gene of bacteriophage T4 have mobile group I introns. The mobility of these introns is also mediated by site-specific endonucleases. However, endonucleases encoded by these introns lack the LAGLI-DADG motifs and share short peptides in common with each other and with a minority of intron ORFs of mitochondrial DNA of filamentous fungi (Michel and Dujon, 1986). A further distinction between these endonucleases and those in yeast is that the bacteriophage intron-encoded enzymes cleave at a distance (up to 25 bp) from the intron insertion site. Thus, it appears that at least two families of intron-encoded endonucleases exist, one represented by yeast HO, the other by td I-TevI. Nevertheless, a comparison of the homing sites for the td and aI4α (HO family) introns shows a 9 out of 14 nucleotides identity (Perlman and Butow, 1989).

In *Chlamydomonas*, group I introns have been described in chloroplast (Cote et al. 1993; Cote and Turmel, 1995; Durrenberger and Rochaix, 1991; Turmel et al. 1995a) and mitochondrial genome (Colleaux et al. 1990). Like group I introns in other organisms, most of the ones found in *Chlamydomonas* also contain ORFs capable of encoding a protein. The deduced protein sequences from the intron-encoded ORFs are significantly similar to the yeast HO family of site-specific endonucleases, as the characteristic LAGLI-DADG motif has been identified in all the ORFs. Among those intron-encoded

ORFs, several of them have been demonstrated to be DNA endonucleases that cut DNA at specific site and leave a 4-nt single-stranded overhangs with 3'-OH termini (Turmel et al. 1995b; Cote et al. 1993; Ma et al. 1992; Thompson et al. 1992; Marshall et al. 1991; Durrenberger and Rochaix, 1991; Gauthier et al. 1991).

(iii). DNA endonucleases are themselves mobile genetic elements

One popular view considers that catalytic RNAs were the primordial elements and that the ORFs were more recently added to the introns (Lambowitz, 1989). The demonstration that foreign sequences within an intron or flanked by exons in the absence of intron sequences were efficiently transferred in an endonuclease-dependent manner, regardless of the presence or absence of intron sequences (Bell-Pedersen et al., 1990), supports the contention that it is indeed the ORF rather than the intron itself that is the mobile progenitor (Perlman and Butow, 1989). In two closely related species of *Neurospora*, both mitochondrial ND1 genes contain group I introns, the intron insertion sites are exactly the same, and the core structure of the introns has 97% identity within a two-hundred-nucleotide sequence, suggesting that the ND1 gene descended from a relatively recent common ancestor. In contrast to this conservation of position and secondary structure, introns in these two strains contain completely different open reading frames located at different positions relative to the core structure element. One is located in frame with the upstream exon, while the other is located downstream of the intron core as a free standing open reading frame. One open reading frame contains homologs of the conserved twelve amino-acid blocks typical of the endonuclease-like open reading frames found in most group I introns, while the other does not (Mota and Collins, 1988). The difference of position and composition between these open reading frames clearly indicate that they were gained independently.

The bacteriophage T4 *segA* gene lies in a genetically unmapped region between two known genes (Sharma et al., 1992). There are also four other genes of unknown function similar to *segA*. They are all similar to another family of endonucleases encoded by group I introns (represented by *td* I-TevI). More significantly, the endonuclease activity of the *segA* gene has been demonstrated in an *in vitro* assay (Sharma et al., 1992). Like the endonuclease encoded by T4 *td* intron, the endonuclease encoded by *segA* is Mg-dependent and site-specific. The finding clearly indicates that endonuclease can insert its coding DNA in the intergenic region independent of intron homing.

D. Protein splicing

Protein splicing is a post-translational processing event in which the expression of a single gene results in production of two proteins. The two products are not from a site-specific protease cleavage. Instead, one of the products is produced from the precise excision of an internal protein region of a precursor, and the two remaining flanking regions are joined together by a *bona fide* peptide bond to form another. Since the sequence of the products of a protein splicing event is not colinear with the mature mRNA, the discovery of protein splicing added another dimension to the flow of genetic information from DNA to RNA to protein (Shub and Goodrich-Blair, 1992). Although there are only a few cases of protein splicing discovered so far, they are represented by three major phyla: yeast *Saccharomyces cerevisiae* (Kane et al., 1990) and *Candida tropicalis* (Gu et al., 1993) in eukaryotes, *Mycobacterium tuberculosis* (Davis et al., 1992) and *Mycobacterium leprae* (Davis et al., 1994) in eubacteria, and *Thermococcus litoralis* (Perler et al., 1992) and *Pyrococcus* species strain GB-D (Xu et al., 1993) in archaea.

(i). Protein splicing in yeast

The first example of protein splicing was described by Kane et al. in 1990. In their laboratory a dominant allele of TFP1 (resistant to the drug trifluoperazine) from a mutant strain of yeast *Saccharomyces cerevisiae* was isolated, and sequencing of the wild-type TFP1 revealed that it contains an ORF encoding a 119-kDa protein. Analyzing the deduced amino acid sequence revealed that it is 73% and 77% identical to a *Neurospora* proton-ATPase 70-kDa subunit at its N- and C-terminus, respectively, with a big stretch of sequence in the middle (counting for 50-kDa of the 119-kDa protein) not similar to any proton-ATPase related protein sequences. Instead, it shows 31% sequence identity to one region of the yeast HO endonuclease. The homology disruption happened to be in a very conserved region that is thought to be part of the ATP-binding site of the TFP1 protein. Analysis of RNA sequence did not show any recognizable characteristics of any known RNA introns, and careful Northern hybridization did not detect any spliced mRNA. RNA introns sometimes contain ORFs, and these intron-encoding ORFs can be in frame with 5' exon, 3' exon, or as a free standing ORF, but there has never been such a big RNA intron in frame with both 5' and 3' exons without a single termination codon within the ORF.

When the ORF was expressed in yeast, a 69-kDa protein was observed (with a tiny amount of 119-kDa precursor). Introduction of a stop codon or in-frame deletion revealed that the continuity of the ORF is necessary for the production of the 69-kDa protein. Western blot with domain-specific monoclonal antibodies suggest the 69-kDa protein contains the N- and C-terminal parts of the 119-kDa protein, while a 50-kDa protein was recognized by antibody specific to the middle part (non-homologous region) of the 119-kDa protein. When the protein was generated *in vitro* using mRNA made via *in vitro* transcription or isolated from yeast, or even when the gene was expressed in *E. coli*, two proteins (a 69-kDa protein and a 50-kDa protein) resulted. Although there were no protein sequencing data confirming that the 69-kDa protein contains the N- and C-terminus of the ORF joined by a peptide bond, all the evidence pointed to the possibility that the 69-kDa protein is the product of protein splicing from the 119-kDa precursor. Since the process can take place in yeast, in *E. coli*, and in rabbit reticulocyte lysate, it is likely that the excision is autocatalytic. To date, there is only one further example of protein splicing described in yeast. In that case, it was found in the homologous gene in *Candida tropicalis*, with the spacer domain located at the same position as in *Saccharomyces cerevisiae*, suggesting that the FTP1 genes in these two species were possibly derived from a recent common ancestor. Interestingly, while the mature proteins share 87.4% overall sequence identity, the middle part has only 31% sequence identity, mostly at the N- and C-termini.

(ii). Protein splicing in archaebacteria

In archaebacteria, the protein goes through protein splicing is the DNA polymerase, which is now widely used in PCR applications (Vent DNA polymerase, NEB). Cloning of the DNA polymerase gene from the archaeon *Thermococcus litoralis* identified an ORF capable of encoding a protein twice the size of a DNA polymerase purified from the same species by other methods. DNA sequence identified two apparent insertional sequences (Perler et al., 1992). Like the situation in yeast TFP1, these two insertional sequences (IS1 and IS2) form a single ORF with their upstream and downstream sequences, dividing the DNA polymerase ORF into three parts. Mutagenesis experiments demonstrated that the intact IS2 ORF is absolutely required for the production of functional DNA polymerase, indicating that the IS2 ORF is translated together with its flanking DNA polymerase sequence domains. Silent mutations at splicing junctions did not affect splicing at all, suggesting that the splicing of IS2 does not occur at the level of RNA. In order to see the relationship between the precursor and the products, pulse-chase analysis was conducted. During the time course of chasing, precursor decreased whereas products increased, thus confirming that the products were indeed produced from the precursor post-translationally. Both IS1 and IS2 are inserted in the same gene, but there is a big difference between them. IS2 can splice in several heterologous systems as the example in yeast *Saccharomyces cerevisiae*, indicating that it is self-splicing, however, IS1 does not splice in any other systems tested. Since the whole gene (DNA polymerase with both IS1 and IS2) is not clonable in *E. coli*, it is not known whether the IS2 can assist the splicing of IS1, or whether other factors in the archaeal cell assist the splicing of IS1, or whether IS1 can self-splice in the archaeal cell. If there are such factors in the archaeal cell assisting the splicing of IS1, what are they?

(iii). Protein splicing in eubacteria

Protein splicing also exists in eubacteria (Davis et al., 1992). When the *Mycobacterium tuberculosis recA* locus, which comprises an 85-kDa ORF, was expressed in *E. coli* (maxicell labeling), two proteins were observed. One was a 38-kDa RecA protein that apparently contains N- and C- termini as judged by Western blot with domain-specific antibodies, the other was a 47-kDa spacer protein, also confirmed by Western blot. Once more, there is no evidence of RNA splicing, as judged by the absence of spliced mRNA in reverse transcription PCR followed by Southern hybridization, which is a very sensitive method for detection of mRNA. In-frame deletion from the middle changed the size of the spacer protein as well as that of the precursor, but not the size of the RecA protein. Deletion from both ends until 8 amino acid residues at the N terminus and 20 amino acid residues at the C terminus of the spacer protein did not affect the production and the size of the spacer protein. These results suggested two things: a, the 47-kDa spacer protein indeed arose from the spacer; and b. the spacer protein probably contains all the information required for its splicing.

Screening other *Mycobacterial* species revealed a protein intron in *Mycobacterium leprae* RecA protein, but other mycobacterial RecAs do not contain such a protein intron (Davis et al., 1994). Unlike the situation with the *Mycobacterium tuberculosis recA* gene, which is spliced to form two smaller proteins even when it is expressed in different genetic systems, the *Mycobacterium leprae* RecA precursor only spliced completely in *Mycobacterium leprae* itself. Changing pH, temperature, oxidative state and ions all failed to detect splicing when it is expressed in *E. coli*. In fact, these two mycobacterial protein introns are different in size, sequence and location of insertion of their coding sequences in RecAs of *Mycobacterium tuberculosis* and *Mycobacterium leprae*, indicating that the acquisition of the protein introns occurred independently in these two species.

So far, protein splicing in eubacteria has only been described in *Mycobacterium tuberculosis* and *Mycobacterium leprae*, and both species of mycobacteria are major human pathogens. Given the rarity of such genetic elements, their presence in the same gene in the two pathogenic species can not be simply explained as that they have arisen by chance, they may have some function for their host organisms. The *recA* gene is important for the repair of DNA damage caused by oxidative stress, one of the conditions these intracellular pathogens are exposed to when they invade macrophages. It is therefore likely that *recA* would be important for survival within the cell and that the splicing of protein introns may be an additional step in the regulation of *recA* expression under certain conditions. It is also possible that they might possess some other functions more directly involved in pathogenesis. In any event, protein introns may possess some function important to the survival of their host rather than just selfish genetic elements.

(iv). Mechanism of protein splicing

Protein splicing can in principle be simply viewed as another kind of intron splicing, as both RNA introns and protein introns are discarded during the process of information transfer from genes to mature products. In parallel with RNA splicing, protein introns have now been named inteins, and sequences flanking inteins are called exteins. Although protein splicing and RNA splicing apparently achieve the same goal by disposing of some information, the structural and chemical mechanisms involved in protein splicing must be fundamentally different from those involved in RNA splicing on the chemical ground.

The alignment of primary sequences of inteins shows very low similarity between any pair of them. Nevertheless, there is at least one thing in common among all the inteins. All the inteins have an N-terminal amino acid residue with a hydroxyl or thiol group at the side chain (Ser, Thr, or Cys) and a His-Asn-Ser/Thr/Cys motif at the C-terminal junction. The universality of the motif and necessity of its function in protein splicing have prompted speculation that this motif may be reminiscent of the catalytic triad found in serine and cysteine proteases (Hodges et al., 1992, Wallace, 1993). These classes of protease proceed via a mechanism where the susceptible peptide bond of the substrate is attacked by a nucleophilic residue (*Ser* or *Cys*), forming an acyl-enzyme intermediate with the N-terminal peptide that is then hydrolyzed by water to release free enzyme and cleaved substrate (Fersht, 1977). The nucleophile is activated for both acylation and deacylation steps by deprotonation and protonation of the side chain by an adjacent histidine, which in turn has its pKa modulated by interactions with an adjacent Asp (in serine proteases) or Asn (in cysteine proteases) residue. Although these residues are not adjacent in the primary sequence of any known proteases, they are in close proximity in the folded active site. Replacement of amino acid residues at splicing junctions of inteins by site-directed mutagenesis results in reduced splicing activity or no splicing at all. Among these changes, replacement of nucleophilic amino acid residues at either junction with other nucleophilic amino acid residues reduced splicing activity dramatically, whereas replacing these residues with non-nucleophilic ones usually abolished the activity totally, strongly suggesting the involvement of the nucleophilic side chain of these residues in protein splicing. More interestingly, the replacement of the Asn residue in the C terminal His-Asn-Ser/Thr/Cys motif with any other residues resulted in the complete loss of splicing activity, suggesting that the Asn residue might play more important roles in protein splicing than it does in proteolysis catalyzed by cysteine proteases. In fact, substitution of either of the splicing junction cysteine residues with glycine in yeast *Saccharomyces cerevisiae* TFP1 gave a protein species corresponding to a precursor that had undergone a cleavage event at the C terminal junction, plus a protein species corresponding to the C terminal extein, as judged by apparent sizes and detection by Western blot with domain-specific antibodies (Cooper et al., 1992). Protein species corresponding to the C-terminal extein has also been observed with wild-ty; TFP1 (Kane et al., 1990). Based on these observations, a model was proposed in which the splicing is initiated by the asparagine residue attacking its C-terminal peptide bond, resulting in cleavage of the peptide bond and the formation of a C-terminal succinimide ring. The ability of an asparagine residue to cause peptide bond cleavage had been demonstrated in peptide and protein cleavage (Clarke et al., 1992, for example).

A big breakthrough in deciphering the mechanisms of protein splicing resulted from the development of an *in vitro* splicing assay (Xu et al., 1993). In the *in vitro* assay, a fusion protein consisting of maltose-binding protein (M: N-extein), the *Pyrococcus* sp pol intein (I: intein), and paramyosin (P: C-extein), was used as the precursor protein. The maltose-binding protein provided a simple purification of the precursor protein, antibodies to different parts of the precursor (M, I, and P) aided in the identification of the precursor protein and the spliced products (including intermediates). Analysis of the MIP fusion expressed in *E. coli* at 12-32 °C revealed that the precursor protein spliced at an extremely slow rate, not surprisingly since *Pyrococcus* intein is an extreme thermophile that grows at temperatures exceeding 95 °C. Purification by amylose affinity chromatography and Mono Q fine protein liquid chromatography provided a relatively pure source of MIP precursor for *in vitro* splicing studies. In the *in vitro* splicing reaction, the conversion of precursor (MIP) to the spliced protein (MP) and the intein (I) products was observed with only the MIP precursor and salts, indicating that the protein splicing reaction is autocatalytic. During the time course of the protein splicing reaction, an intermediate species was detected that migrated at a significantly slower rate on SDS-PAGE than the precursor. Western blot analysis demonstrated that it contains the intein and both exteins, and protein sequencing revealed two sequences corresponding to the amino termini of both N-extein and intein. This slow-migrating protein species was thus demonstrated to be a branched intermediate, which is a surprising discovery in that it is reminiscent of the branched RNA intermediates (lariats) formed during group II intron and nuclear mRNA intron splicing.

Two remarkable observations from the *in vitro* protein splicing reaction are important to the consideration of models for the splicing mechanisms. One is that the formation of the branched intermediate is reversible, the other is the presence of a succinimide ring in a carboxyl terminal peptide isolated from the spliced intein, providing solid evidence that at some step in the protein splicing reaction the invariant intein carboxyl terminal Asn undergoes a succinimide rearrangement. Based on the experimental data that have accumulated, several models for the protein splicing mechanism have been proposed (Wallace, 1993, Xu et al., 1994, Clarke, 1994). A big challenge for these models is whether the splicing mechanism for the inteins of extreme thermophiles reflects the splicing mechanism for inteins in nonthermophiles and inteins with Cys residues at their splicing junctions. The observation of other molecular species in yeast *Candida tropicalis* (N-terminal extein, Gu et al., 1993) and in eubacteria *Mycobacterium tuberculosis* (N-terminal extein-intein and intein-C-terminal extein, Davis et al., 1992), both of which are nonthermophiles and have Cys at their splicing junctions, leaves room for further modifications (in my view) of the protein splicing models proposed.

(v). Inteins are DNA endonucleases

surprisingly, another factor common to all the intein sequences is the presence of two dodecapeptide sequence motifs corresponding to the active sites of a family of DNA endonucleases (represented by yeast HO endonuclease). Dodecapeptide sequences are present in endonucleases encoded by open reading frames within group I self-splicing introns that exhibit genetic mobility. The endonuclease mediates the unidirectional transfer of the intron from a copy of the gene that contains the intron to one that lacks it, a process termed intron homing. So far four of the protein splicing inteins have been shown to be site-specific endonucleases: the *Saccharomyces cerevisiae* VMA intein, the *Thermococcus litoralis* DNA polymerase inteins I and II, and the *Pyrococcus sp pol* intein (Perler et al., 1994). Moreover, not only has the *Saccharomyces cerevisiae* VMA intein been shown to be a site-specific endonuclease, it also specifically cleaves at the site of insertion of the intein in a copy of the gene that lacks it and converts the intein-less gene into an intein-containing gene (Gimble and Thorner, 1992). The mechanism of mobility for inteins is the same as that for group I self-splicing introns, as both are intervening sequence elements that encode an endonuclease able to mediate genetic mobility, and both remain phenotypically silent to the host by either RNA or protein splicing.

While only four of the identified inteins have been demonstrated to be site-specific endonucleases, all the identified inteins are at least derived from endonucleases because all the inteins have the dodecapeptide sequences found in homing endonucleases and yeast HO endonuclease (Pietrokovski, 1995). Interestingly, mutagenesis of the dodecapeptide sequence of *Thermococcus litoralis* DNA polymerase intein II inactivated the intein endonuclease activity, but the protein splicing activity was still retained, demonstrating that the two distinctive activities are not associated. The genetic mobility of at least one intein (yeast VMA intein) coupled with the extremely diverse phylogenetic distribution of inteins (Eukaryotes, bacteria and archaea) suggests a horizontal mode of transmission. This is most clear in the case of *Mycobacterium recA* inteins, where inteins were found only in two of the 33 *Mycobacterium* species investigated so far, and they are different in size and insertion points within the *recA* coding region, suggesting that they arose from two independent insertion events.

Chapter II. Materials and Methods

MATERIALS

General chemicals were from BDH Inc., Anachemia, Boehringer Mannheim Biochemicals and Sigma Chemical Company. Tris, polypeptone and yeast extract were from Bethesda Research Laboratories (BRL). T7 DNA ligase, T7 RNA polymerase, T7 DNA polymerase, RNase-free DNase I, exonuclease III, S1 nuclease and Klenow fragment of *E. coli* DNA polymerase I were from BRL. IPTG and X-gal were from ICN. Restriction endonucleases were from BRL, New England Biolabs (NEB) and Promega. M-MLV reverse transcriptase was from Pharmacia. Sequenase and sequencing kit were from United States Biochemicals Corp. Protein markers were from Promega and Sigma. Taq DNA polymerase was from BRL, Vent DNA polymerase was from NEB. [α-^{32}P]-dATP and [α-^{35}S]-dATP were from DuPont, NEN. pET vector plasmid was from Novagene, pMAL vector plasmid was from NEB. Clones BS-7 and S-7, which contain DNA fragments from *Guillardia theta* chloroplast genome, and *Guillardia theta* total RNA were from Dr. Susan E. Douglas, Institute for Marine Biosciences, Halifax, Nova Scotia, Canada. Total DNA of various *Chlamydomonas* species was from Dr. Robert R. Lee, Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada. Membranes used in Southern, Northern and Western blotting experiments were from Micron Separations Inc. (MSI), PVDF membrane was from BioRad. Peroxidase-conjugated goat anti-rabbit IgG antibody was from BRL, LumiGlo Chemiluminascent substrates A and B were from KPL Inc. Hyperfilm-ECL was from Amersham, X-ray film was from Eastman Kodak. Oligonucleotides used in PCR-mediated site-directed mutagenesis were from Dalton Chemical Laboratories Inc.

METHODS

(i). Methods involved in DNA manipulation

a. Preparation of DNA

1. Large-scale preparation of plasmid DNA (alkaline lysis)

A single colony was grown in 100 ml of LB medium (1% tryptone, 0.5% yeast extract, 1% NaCl, pH 7.0) plus antibiotics overnight at 37 °C with constant shaking. Cells were harvested by centrifugation, resuspended in 6 ml of solution I (50 mM glucose, 10 mM EDTA, 25 mM Tris-HCl, pH 8.0, and 3 mg/ml lysozyme). After 5 min at room temperature, 12 ml of solution II (0.2 M NaOH, 1% SDS) was added and the solution mixed quickly by inverting, then left on ice for 5 min. Nine ml solution III (5 M KOAc, pH 5.0) was added and the solution mixed by inverting and left on ice for 5 min. The mixture was centrifuged at 20,000 Xg for 20 min at 4 °C, and the supernatant was filtrated into a new tube. DNA in the supernatant was precipitated by the addition of isopropanol and centrifuged at 10,000 Xg for 10 min at room temperature. The DNA pellet was washed with 70% ethanol, air dried, and resuspended in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0).

2. Mini-preparation of plasmid DNA (boiling method)

A single colony of bacterial cells was inoculated into 2 ml LB containing antibiotics and grown at 37 °C overnight. Cells were harvested as a pellet in an Eppendorf tube and resuspended in the residual liquid remaining in the tube by vortexing briefly. 350 µl of STET buffer (8% sucrose, 5% Triton X-100, 50 mM EDTA, 50 mM Tris-HCl, pH 8.0) plus 25 µl of 10 mg/ml lysozyme was added and mixed well and left at room temperature for 5 min. The mixture was heated in a boiling water bath for 40 sec and the lysate centrifuged for 10 min at 14,000 rpm to remove the fluffy pellet. DNA was precipitated by addition of 350 μl isopropanol, left at room temperature for 5 min, centrifuged for 5 min. The DNA pellet was washed with 70% ethanol, air dried and resuspended in 30 μl of TE buffer.

3. Purification of plasmid DNA on CsCl gradient

A single colony of *E. coli* cells harboring the plasmid of interest was grown in 100 to 200 ml of 2x YT medium (1.6% Tryptone, 1% yeast extract, 0.5% NaCl, pH 7.0) containing antibiotics (usually ampicillin) at 37 °C overnight. Cells were harvested and plasmid DNA was prepared using the large-scale alkaline lysis method as described in (i).-a.-1; above. At the final step, the DNA pellet was dissolved in 700 μl TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). 1.225 g of CsCl was added and dissolved by mixing, and the resulting solution was centrifuged for 10 min to pellet the aggregated materials. The supernatant was transferred to a new tube and mixed with 90 μl of 10 mg/ml EtBr, then centrifuged for 10 min at room temperature to pellet the aggregated materials. The supernatant was transferred to an ultracentrifuge tube, balanced with 43% CsCl, and centrifuged at 100,000 rpm for 2.5 hrs at room temperature (TL100, Beckman). Supercoiled plasmid band was collected in a dark room in front of a long wavelength UV box. DNA was extracted with water-saturated isoamyl alcohol several times until the isoamyl alcohol was colorless. DNA was diluted with 3 volumes of water, precipitated with ethanol, then recovered by centrifugation at 10,000 Xg for 20 min, air dried and resuspended in TE buffer.

b. DNA digestion, fragment purification and ligation to vectors

DNAs or plasmid vectors were digested with various restriction endonucleases in the buffers recommended by enzyme suppliers. The resultant fragments were resolved by electrophoresis in a 0.8% agarose gel buffered by TAE (40 mM Tris-base, 2 mM EDTA, 20 mM NaOAc, 29.6 mM HOAc, pH 7.8). The gel was stained with EtBr, DNA bands were visualized under long wavelength UV light and were excised. The gel pieces containing insert DNA bands were incubated in 3 volumes of 6 M NaI at 50-55 °C for 5-10 min until the gel was completely dissolved. The solution was chilled on ice and mixed with GlassMilk (Bio101) and left on ice for 5 min. The GlassMilk was centrifuged for 5 sec and pellet was washed three times with NEW buffer (50% ethanol, 100 mM NaCl, 1 mM EDTA, 20 mM Tris base), DNA was eluted by extraction twice with 10 µl of water by incubation at 45-50 °C for 5 min. For blunt end ligation, insert DNA and vector DNA were incubated with dNTPs and Klenow DNA polymerase in ligase buffer for 5 min at room temperature prior to the addition of ligase; for sticky end ligation, this step was omitted. The ligation reaction proceeded at 16 °C for 5 hrs or 4 °C overnight.

c. Transformation of plasmid DNA into *E. coli* cells

1. Preparation of competent *E. coli* cells by the DMSO method

For preparation of competent *E. coli* cells, 1 ml of overnight culture was diluted into 100 ml LB medium ((1% peptone, 0.5% yeast extract, 1% NaCl (w/v)), and grown at 37 °C until the A_{600} was 0.3-0.4, after which the culture was cooled on ice. Cells were collected by centrifugation for 5 min at 4 °C at 1000 Xg and resuspended in 10 ml of ice-cold TSB buffer (LB medium, pH 6.1, 10% PEG 3350, 5% DMSO, 10 mM $MgCl_2$, 10 mM $MgSO_4$). Cells were aliquoted and quickly frozen in a -70 °C ethanol bath. Competent cells were stored at -70 °C until use.

2. Preparation of competent *E. coli* cells by the calcium chloride method

One ml of overnight culture grown from a single colony was diluted into 100 ml LB medium and grown at 37 °C for 2-3 hrs until the A_{600} was 0.3-0.4. Cells were harvested and resuspended in 20 ml of 0.1 M $CaCl_2$, then left on ice for 5 min. Cells were recovered by centrifugation at 4000 Xg for 10 min at 4 °C and resuspended in 4 ml $CaCl_2$. Cells were aliquoted and quickly frozen in a -70 °C ethanol bath. Competent cells were stored at -70 °C until use.

3. Transformation of *E. coli* cells with plasmids

For transformation of *E. coli* cell prepared by the DMSO method, 10 µl of ligated DNA was mixed with 100 µl competent cells and left on ice for 15 min. TSB buffer (900 µl) containing 20 mM glucose was added and the mixture incubated at 37 °C for 30 min. A 200 µl culture was plated onto an LB plate containing ampicillin (50 µg/ml) and was incubated at 37 °C overnight. For transformation of *E. coli* cells prepared by the calcium chloride method, 10 µl of ligated DNA was mixed with 200 µl of competent cells and left on ice for 30 min. The mixture was then heat-shocked at 42 °C for 100 sec, mixed with 800 µl of LB medium without antibiotics, and incubated at 37 °C for 45 min. An aliquot of 200 µl culture was plated onto an LB plate containing ampicillin (50 µg/ml) and incubated at 37 °C overnight.

(ii). DNA sequencing

a. Systematic deletion

In general, 5-10 µg of closed circular plasmid DNA was cut with a restriction endonuclease such as Pst I or Sst I, which will generate a 4 bases 3'-protruding end, and another restriction endonuclease such as EcoR I or BamH I, which will generate a 5'-protruding end. The 3'-protruding end is closer to the vector, and is exonuclease III (Exo III) resistant, whereas the 5'-protruding end is closer to the insert and is Exo III digestible. DNA was extracted with phenol/chloroform/iso-amyl alcohol (25/24/1), precipitated with ethano! and dissolved in 60 µl Exo III buffer (66 mM Tris-HCl, pH 8.0, 0.66 mM $MgCl_2$), 250-500 units of Exo III were added and 2.5 µl samples were removed at 30 sec intervals and mixed with 7.5 µl S1 mix [27 µl 7.4x S1 buffer (0.3 M KOAc, pH 4.6, 2.5 M NaCl, 10 mM $ZnSO_4$, 50% glycerol), 172 µl H_2O, 60 units S1 nuclease]. Samples taken from each time point were incubated at room temperature for 30 min, then 1 µl of S1 stop buffer (0.3 M Tris base, 0.05 M EDTA) was added and the reaction mixture heated at 70 °C for 10 min. Progressive unidirectional deletion was examined by taking 2 µl from each time point and electrophoresing on 1% agarose gel. To each time point tubes 1 µl of Klenow mix [30 µl 1x Klenow buffer (20 mM Tris-HCl, pH 8.0, 100 mM $MgCl_2$), 3-5 units of Klenow fragment of *E. coli* DNA polymerase I] was added and incubated at 37 °C for 3 min. then 1 µl dNTP mix (0.125 mM each of dATP, dCTP, dGTP, dTTP) was added and incubated at 37 °C for another 5 min. Samples were removed to room temperature and 40 µl of ligase mix [790 µl H_2O, 100 µl 10x ligase buffer (500 mM Tris-HCl, pH 7.6, 100 mM $MgCl_2$, 10 mM ATP), 100 µl PEG 8000, 10 µl 100 mM DTT, 5 units T4 DNA ligase] were added and the mixture incubated at room temperature for 1 hr. Ten µl samples were used to transform *E. coli* cells.

b. Preparation of ssDNA for sequencing

Twenty µl of overnight culture grown from a single colony were inoculated into 2 ml of 2x YT (1.6% Tryptone, 1% yeast extract, 0.5% NaCl, pH7.0) medium containing ampicillin and grown at 37 °C for 1.5 hrs. One hundred µl of 2x YT containing 2 µl of helper phage (M13 KO7), 1 ml of 2x YT containing 50 µg/ml of ampicillin, 210 µg/ml of kanamycin were added and grown at 37 °C for 14-18 hrs. An aliquot (1.2 ml) was taken from 1.5 ml of culture, centrifuged for 5 min, and 300 µl of PEG (20% PEG8000, 2.5 M NaCl) was added to the supernatant, which was then mixed by inverting and left at room temperature for 15 min. Phage particles were collected by centrifugation at room temperature for 10 min and resuspended in 100 μl of TES (TE plus 100 mM NaCl). DNA was extracted once with 50 μl phenol, once with 50 μl CIA (chloroform/iso-amyl alcohol, 24/1). DNA was precipitated with 2.5 volumes of ethanol and washed once with 70% ethanol, and the pellet was air-dried. DNA was resuspended in 20 μl of TE buffer. The size and concentration of each ssDNA was estimated by electrophoresis of 1 μl of ssDNA on a 0.8% agarose gel.

c. Preparation of dsDNA for sequencing

The preparation of dsDNA was essentially the same as described for large-scale preparation of plasmid DNA. For dsDNA sequencing, 1 μg of DNA in 60 μl of TE was mixed with 6 μl of 2 N NaOH, left at room temperature for 5 min, then 24 μl of 5 M ammonium acetate (pH 7.4) and 3 volumes of ethanol was added. DNA was pelleted by centrifugation, washed once with 70% ethanol, and air dried. Denatured DNA was stored as a pellet until use.

d. DNA sequencing by the dideoxy-chain termination method

DNA sequencing was performed by the dideoxy chain-termination method (Sanger, 1977), using Sequenase Version II according to the supplier's protocol (United States Biochemicals Corp.). Both strands of the DNA were sequenced. Most sequences were from ssDNA sequencing, with small gaps being filled with dsDNA sequencing using synthetic oligonucleotides as primers and dsDNA purified from corresponding subclones as templates. Sequencing reaction products were resolved in 6% polyacrylamide, 7 M urea gel buffered with TBE (50 mM Tris, 50 mM boric acid, 1 mM EDTA, pH 8.3) by electrophoresis for 2.5 h at 2000V. The gel was soaked in 10% acetic acid, 10% methanol to remove the urea, then dried onto a 3MM paper (Whatman) using a

BioRad slab gel dryer before expose to X-ray film overnight. DNA sequences were analyzed using MicroGenie™ (Queen and Corn, 1984) and DNA Strider™ programs (Marck, 1988), and the deduced amino acid sequences were aligned by using sequence alignment program of Corpet (1988).

(iii). Hybridization

a. Northern Blot

A 1.1% agarose gel was used to resolve the *Guillardia theta* total RNA (from Susan E. Douglas). In gel preparation, 0.88 g agarose was mixed with 50 ml water, melted and cooled to 60-65 °C. In the fume hood, 14 ml of formaldehyde and 16 ml of 5x gel buffer (0.1 M MOPS, 40 mM NaOAc, 5 mM EDTA, adjust pH to 7.0 with NaOH) were added and the solution swirled to mix. The gel was poured and left at room temperature for at least one hr. For sample preparation, 5 µg *Guillardia theta* total RNA in 4.5 µl water was mixed with 2.0 µl of 5x gel buffer, 3.5 µl of formaldehyde and 10 µl of formamide. The mixture was incubated at 70 °C for 10 min., chilled on ice, and then 2 µl of loading buffer (50% glycerol, 1 mM EDTA, 0.4% bromophenol blue) were added. The gel was prerun at 60 V for 3 min in 1x gel buffer, with the gel never being submerged with the gel buffer. Samples were loaded into the dry well and the gel was run at 60 V for 5 min to let the sample enter the gel, then run at 100 V for 1 hr. The gel was taken out of the tank, mixed with the electrophoresis buffer, put back, and continued to run for another one hr until the dye front reached two-thirds of the gel length.

The RNA marker lane together with one of the RNA sample lanes was cut and stained in 0.5 µg/ml of EtBr for 15 min and destained in water for 30 min or longer until RNA bands were clearly visible under UV light. Migration distances of each of the RNA bands in both lanes were recorded. The rest of gel, which contains the total RNA of *Guillardia theta*, was washed in 1 liter of water for 5 min., the wash was repeated for 6 times and the gel equilibrated in 20x SSC (3 M NaCl, 0.3 M Na-citrate, pH 8.0) for 10 min. The gel was placed on a glass plate covered with two layer of 3 MM paper with both ends in 20x SSC. The gel was covered with a piece of 0.22 micron Nylon membrane and two pieces of filter paper the same size as the gel. Filter papers and nylon membranes were soaked in 20x SSC before use. On the top, 6-10 inches of dry paper towel was applied and left overnight. Next morning the membrane was removed, placed on a piece of dry filter paper with the RNA side facing up, and air dried for 30 min. RNA was fixed onto the membrane by UV-cross linking and used in subsequent hybridization.

b. Southern blot

For Southern blot experiments, DNA was resolved in 0.8% agarose gel buffered with TBE by electrophoresis. Electrophoresis was stopped when the dye front reached two-thirds of the gel length. The gel was stained with EtBr and the migration distances of DNA molecular weight markers (Lambda DNA Hind III fragments, BRL) were recorded. The gel was soaked in denaturing buffer (1.5 M NaCl, 0.5 M NaOH) for 45 min, then in neutralizing buffer (1 M Tris-HCl, pH 7.4, 1.5 M NaCl) for 30 min. Transfer to the Nylon membrane and the subsequent UV cross link were essentially the same as described for Northern blot analysis.

c. Probe labeling

DNA probes were labeled using Prime It Random Primer Kit (Strategene) with [α-^{32}P]-dATP (DuPont). In general, 25 ng probe DNA in 24 µl TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) was mixed with 10 µl random primer (9-mer random oligos, 27 OD/ml), heated at 100 °C for 5 min, chilled on ice, then 10 µl of 5x labeling buffer (200 mM Tris-HCl pH 7.5, 50 mM $MgCl_2$, 5 mM DTT, 0.1 mM dCTP, dGTP and dTTP), 5 µl of [α-^{32}P]-dATP (3000 Ci/mmole, 10 µCi/µl), and 1 µl T7 DNA polymerase (2u/µl) were added. The reaction mixture was incubated at 37 °C for 5 min, then 2 µl of 0.5 M EDTA, 130 µl of water, and 20 µl of 2.5 NaAc (pH 5.0) were added. DNA was precipitated by the addition of 2 volumes of ethanol, stored at -70 °C for 30 min, and centrifuged for 5 min in a benchtop centrifuge. DNA was resuspended in 100 µl of hybridization buffer (5X Denhardt's, 6X SSPE, 1% SDS, 100 µg/ml denatured salmon sperm DNA).

d. Hybridization

One strip of membrane bearing total *Guillardia theta* RNA or DNA was added to 10 ml of hybridization buffer, sealed into a plastic bag, and incubated at 68 °C for 2 hrs with slow shaking. The labeled probe was heated at 100 °C for 5 min and added to the bag. The incubation was continued at 68 °C overnight with slow shaking. The strip was washed with washing buffer I (2xSSC, 1% SDS) once at room temperature, twice at 68 °C, and with washing buffer II (0.2x SSC, 1% SDS) once at 68 °C. Signal was visualized by exposure to an X-ray film overnight.

(iv). Polymerase Chain Reaction (PCR)

a. Isolation of total RNA from *E. coli*

To investigate the transcription of *Chlamydomonas eugametos clpP* gene in *E. coli*, total RNA was isolated from *E. coli* cells containing the *clpP* gene bearing plasmid. For RNA isolation, 12 ml of *E. coli* culture were grown at 37 °C until A_{600} was 0.5, IPTG was added at this point to the final concentration of 2 mM, and growth was continued at 37 °C for 30 min. Cells were harvested and resuspended in 400 µl of 2% SDS, then heated at 60 °C for 5 min. Ten ml of trizol (BRL) was added to the tube and vortexed briefly, then 2 ml of chloroform were added and vortexed, the mixture was centrifuged at 5,000 rpm for 10 min. The aqueous layer was transferred to a new tube and 0.6 volume of isopropanol was added to precipitate RNA. The RNA pellet was washed with 70% ethanol and air dried. Total RNA was resuspended in water and digested with RNase-free DNase (BRL) in the presence of RNAGuard (BRL). RNA was extracted with phenol/CIA (phenol/chloroform/iso-amyl alcohol, 25:24:1) once and precipitated with 3 volumes of ethanol. The pellet was washed once with 70% ethanol and air dried. RNA was resuspended with 50 µl of water and the A_{260} was measured to estimate the RNA concentration.

b. cDNA synthesis

For reverse transcription, 5 µg of RNA in 11.5 µl of water were mixed with 5 µl of oligo O-104 (100 ng/µl), heated at 80 °C for 5 min, briefly centrifuged and put into a 55 °C water bath, and then mixed with 6 µl of 5x RT buffer (50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM $MgCl_2$), 1.5 µl of 40 mM dNTPs (10 mM each of dATP, dCTP, dGTP, dTTP), 3 µl of 100 mM DTT, 1 µl of RNAGuard, and 2 µl of M-MLV reverse transcriptase (200 U/µl) or 2 µl of water as control. The mixture was incubation at 55 °C for 1 hr followed by heating at 80 °C for 10 min and then centrifugation at room temperature for 5 min at 14,000 rpm. Two µl of cDNA were used in subsequent PCR.

c. PCR with Vent DNA polymerase

Chlamydomonas reinhardtii chloroplast-encoded *clpP* and *rps3* gene were amplified by Vent DNA polymerase (New England Biolab.). Oligonucleotides (oligos) for the *clpP* gene were O-72: 5'-GGGAATTCATGCCGATTGGAGTACC-3' and O-67: 5'-GGGAATTCCATATTACTAA-3'. Oligos for the *rps3* gene were O-65: 5'-CCCCATGGGTCAAAAGTACA-3' and O-64: 5'-AACTAATCTTAGAGCGT-3'. Generally, in a 100 μl reaction, 0.2 μg of total *Chlamydomonas reinhardtii* DNA was mixed with 10 μl of 10x Vent DNA polymerase buffer (supplied by New England Biolab.) and 5 μl of each oligo (100 ng/μl), heated at 100 °C for 3 min, cooled on ice, with subsequent addition of 10 μl of dNTP (2 mM each of dATP, dCTP, dGTP, dTTP), 1 μl of BSA (100 μg/μl), and 1 μl Vent DNA polymerase (2 U/μl). PCR reaction was performed by incubating samples at 94 °C for 3 min to denature the DNA, then repeating the cycle of 48 °C for 1 min, 62 °C for 10 sec, 72 °C for 3 min, and 94 °C for 30 sec for 35 times. The last cycle was 48 °C for 1 min, 62 °C for 10 sec, and 72 °C for 10 min.

d. PCR with Taq and Vent DNA polymerase

The *Chlamydomonas eugametos* chloroplast-encoded *clpP* gene was amplified by a mixture of Taq and Vent DNA polymerases (5 units of Taq and 0.05 unit of Vent in a 100 μl reaction) with oligos O-105: 5'-CGCCATGGCTATTGGTGTTCCACGTA-3' and O-104: 5'-CCGCTCKAGATTTTTAATTTTGAGATTGTGTAT-3'. PCR was performed based on the protocol described by Watkins et al. (1993). In general, 0.2 μg of total DNA was mixed with 10x buffer (250 mM Tris-HCl, pH 8.5, 160 mM ammonium sulfate, 35 mM MgCl) and 5 μl of each oligos (100 ng/μl), heated at 100 °C for 3 min, cooled on ice, with subsequent addition of 10 μl of dNTP (2 mM each of dATP, dCTP, dGTP, dTTP). In each reaction 5 units of Taq DNA polymerase (BRL) and 0.05 unit of Vent polymerase

(NEN) were used to maximize the yield while minimizing the possibility of non-specific mutation. PCR reaction conditions were 94 ^{0}C for 30 sec to denature the DNA, then 46 ^{0}C for 1 min, 62 ^{0}C for 10 sec, and 72 ^{0}C for 5 min for 35 times, and finally 46 ^{0}C for 1 min, 62 ^{0}C for 10 sec, and 72 ^{0}C for 10 min. PCR amplification of cDNA was carried out under essentially the same conditions as the amplification of *Chlamydomonas eugametos* chloroplast-encoded *clpP* gene described here, except that only Taq DNA polymerase was used in cDNA amplification.

(v). SDS-polyacrylamide gels for proteins

a. Preparative gel (SDS-PAGE)

Resolving gel:

10 ml water, 6 ml 60% sucrose, 8 ml 5x lower gel buffer (2.12 M Tris-HCl, pH 9.18), 16 ml acrylamide/bis-acrylamide (30/0.8), 400 µl 10% SDS, 200 µl 10% ammonium persulfate, 15 µl TEMED.

Stacking gel:

5.3 ml water, 2.5 ml 4x stacking gel buffer (216 mM Tris-H_2SO_4, pH 6.1), 2 ml acrylamide/bis-acrylamide (30/0.8), 100 µl 10% SDS, 100 µl 10% ammonium persulfate, 5 µl TEMED.

Upper tank buffer (10x):

0.82 M Tris-boric acid, pH 8.64, 1% SDS

Lower tank buffer (5x):

2.12 M Tris-HCl, pH 9.18.

b. Analytical gel (SDS-PAGE)

Resolving gel:

3.35 ml water, 2.5 ml 1.5 M Tris-HCl, pH 8.8, 100 μl 10% SDS, 4.0 ml acrylamide/bis-acrylamide (30/0.8), 100 μl 10% ammonium persulfate (fresh weekly), 5 μl TEMED.

Stacking gel:

6.1 ml water, 2.5 ml 0.5 M Tris-HCl, pH 6.8, 100 μl 10% SDS, 1.3 ml acrylamide/bis-acrylamide (30/0.8), 50 μl 10% ammonium persulfate (fresh weekly), 10 μl TEMED.

Upper tank buffer (5x):

125 mM Tris base, 125 mM glycine, pH 8.3, 0.5% SDS.

Lower tank buffer (4x):

1.5 M Tris-HCl, pH 8.8.

c. Protein transfer buffer:

25 mM Tris base, 25 mM glycine, 0.1% SDS, 10% methanol.

(vi). Isolation of 30S ribosome subunits from *Chlamydomonas reinhardtii* chloroplasts

Chlamydomonas reinhardtii cells were grown to 5-8 x 10^6 cells/ml and harvested by centrifugation at 5,000 rpm (JA-10, Beckman) at room temperature for 5 min. The pellet was washed once with SA buffer (25 mM Tris-HCl, 5 mM MgCl, 100 mM KCl, pH 7.8), and resuspended in SA buffer to 4 x 10^9 cells/ml. The cell resuspension was passed through FrenchPress at 5,000 psi to break the cells and the lysate was centrifuged at 18,000 rpm (JA-20, Beckman) at 4 °C to pellet the cell debris and other insoluble materials. Two ml of supernatant were loaded onto 38 ml of SA buffered 10-30% sucrose gradient, centrifuged at 22,500 rpm at 4 °C for 20 hrs, and fractionated by monitoring absorption at 280 nm. The 30S subunit peak was collected and diluted with an equal volume of SA buffer, then loaded onto an SA-buffered 30% sucrose cushion and centrifuged at 48,000 rpm for 22 hrs to pellet the subunits. The pellet was resuspended in water and the protein concentration was measured at A_{280nm}. One unit of absorbance at 280 nm was considered to represent a protein concentration of 1mg/ml.

(vii). Protein purification

a. Protein expression

For protein expression in *E. coli*, a single colony was grown overnight at 37 °C and diluted 100 times with fresh LB medium containing 50 µg/ml ampicillin. The culture was grown at 37 °C for 3 hrs until the A_{600} was 0.5. IPTG was added at this point to the final concentration of 1 mM and cells were continually grown for 2-3 hrs. Cells from 1 ml of culture were harvested by centrifugation at 15,000 Xg for 2 min and the cell pellet was resuspended i. 100 µl of protein loading buffer (1% SDS, 6% sucrose, 1 x stacking buffer, 50 mM DTT, 0.05% bromophenol blue).

b. Protein purification by affinity column chromatography

Fusion protein expressed from pMAL vector was purified by amylose column chromatography. In brief, proteins were induced with the addition of IPTG and the induction was checked by analytical SDS-PAGE. Cells from 100 ml culture were harvested by centrifugation and resuspended in 10 ml of column buffer (20 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM EDTA, 1 mM DTT, 1 mM sodium azide) and sonicated for 2 min. Supernatant was obtained by centrifugation of samples at 9,000 Xg for 30 min (crude extract). The crude extract was diluted 5 times with column buffer and run through an amylose resin affinity column. The column was washed with the column buffer three times. Fusion protein was eluted with column buffer containing 10 mM maltose. Protein concentration was estimated by measuring the absorbance at 280 nm.

c. Protein purification by electroelution

Proteins were induced with the addition of IPTG. Total cellular proteins were resolved by electrophoresis in a preparative gel, the gel was stained (0.1% Coomassie blue R-250, 40% methanol) and destained (50% methanol). Proteins of interest were excised and electroeluted in TGS buffer (25 mM Tris, 192 mM glycine, 0.5% SDS). Protein concentration was estimated by comparison to the protein markers on an analytical SDS-PAGE.

d. Factor Xa digestion of purified fusion protein

Protein for Factor Xa digestion was purified slightly differently. In brief, a single colony was grown at 37 °C overnight, 1 ml of the overnight culture was diluted into 100 ml of fresh LB medium and grown at 37 °C for 2-3 hrs until A_{600} was 0.5. IPTG was added at this point to a final concentration of 1 mM and cells were grown for 2 more hrs. Cells were harvested by centrifugation and resuspended in 10 ml of column buffer (20 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM EDTA, 1 mM DTT, 1 mM sodium azide) and sonicated for 2 min. The insoluble material was obtained by centrifugation at 10,000 Xg for 30 min at 4 °C and the supernatant was discarded. The protein of interest being in the pellet. The pellet was dissolved in 5 ml of protein loading buffer (1% SDS, 6% sucrose, 1 x stacking buffer, 50 mM DTT, 0.05% bromophenol blue) and heated for 5 min in a boiling water bath. Proteins were resolved in SDS-polyacrylamide gel, the gel was stained (0.1% Coomassie, 40% methanol) and destained (50% methanol), and the protein band was excised and protein was electroeluted as described previously. The electroeluted protein was dialyzed against Factor Xa buffer (20 mM Tris-HCl pH 8.0, 100 mM NaCl, 2 mM $CaCl_2$) with three changes of dialyzing buffer at 1 hr, 4 hrs and overnight. Ten μg of protein were incubated with 1 μg of Factor Xa at 23 °C for 2 hrs, the reaction was stopped by the addition of an equal volume of 2 x protein loading buffer and the solution was heated for 5 min in a boiling water bath.

e. Transblotting of proteins from SDS polyacrylamide gels to NitroPlus or PVDF membranes

e.1. Transfer of proteins onto NitroPlus membrane for Western blot or antibody purification

Protein antigens, either expressed in *E. coli* or purified from *Chlamydomonas reinhardtii* chloroplasts, were resolved by SDS-PAGE. After electrophoresis, the gel was covered with a piece of NitroPlus membrane of the same size, then covered on both sides with three pieces of 3MM paper (Whatman). Both membrane and 3 MM paper were soaked with transfer buffer (25 mM Tris, 192 mM glycine, 10% methanol) before use. The gel sandwich was placed between two pieces of sponge and the supporting frame, vertically inserted into a tank filled with transfer buffer, and electrotransferred at 400 mA for 8 hrs or 250 mA overnight After transfer, the membrane was submerged with block buffer (5% skimmed milk powder, 0.2% Tween-20TM, 0.02% NaN_3, 1x PBS) and slowly shaken at room temperature for at least 1 hr before being ready for Western blot.

e.2. Transblotting of proteins to PVDF membrane for protein micro-sequencing

Proteins cut by Factor Xa were resolved in a preparative SDS-polyacrylamide gel and transferred onto PVDF membrane. Generally, the PVDF membrane was soaked in methanol for 1 min, then in transfer buffer (25 mM Tris, 192 mM glycine, 10% methanol) for 5 min before use. Conditions of electrotransfer were the same as those for transfer to NitroPlus membrane for Western blot. After transfer, the membrane was washed in water for 5 min, then stained in Ponceau S staining solution (0.2% Ponceau S, 1% HOAc in water) for 1 min and destained in water with several changes. The area of

PVDF membrane bearing the protein of interest was cut out and sent for protein micro-sequencing.

(viii). Antibodies and Western blot

a. Raising antibodies

Proteins purified by either affinity column chromatography or electroelution were used as antigens to raise antibodies from rabbits. For each injection, 100 µg of proteins were mixed with an equal volume of Freund's adjuvant (Sigma) and injected into rabbits under the back skin, with the antigens being equally distributed into 5-6 injection spots. The injection was repeated at each six weeks and 100 µl blood were collected 10-14 days after each injection. The blood was clotted for 30-60 min at 37 °C, the clot was separated from the wall of a container, and the container was placed at 4 °C overnight. The sera were separated from the clot by centrifugation at 10,000 Xg for 10 min at 4 °C. Antibody activities were tested by Western blot analysis as described below.

b. Antibody purification

Proteins purified by either of the methods described previously (see section vii) were far from pure. When proteins are electroeluted from SDS-PAGE gel for antibody production, non-specific proteins co-migrating with the antigen will not be separated from the antigen. When proteins are eluted from an affinity column, multiple bands were seen after analytical SDS-PAGE and Coomassie staining. Thus non-specific antibodies were raised when such protein preparations were used as antigens. Non-specific antibodies may also arise from unknown sources during the life-time of a rabbit. In order to purify antibodies from the antisera, the following strategy was used. Briefly, one piece of DNA encoding an antigen was cloned into two vectors so the same antigen was expressed as two proteins with different molecular sizes. The two proteins were blotted onto NitroPlus membrane (MSI transfer membrane), and cut out as strip *a* and *b*. Non-specific antibodies that absorb to strip *a* will therefore not absorb to strip *b*. For antibody purification, strip *a* was incubated with antisera at room temperature for 6 hrs with shaking, the strip was then washed with 0.15 M NaCl at room temperature for 20 min, and washed with 1 x PBS at room temperature for 20 min. Antibody was eluted by incubation with 0.2 M glycine (pH 2.8) at room temperature for 20 min. The eluted antibody was then purified by incubating against strip *b*. The purification method was the same as for strip *a*.

c. Western blot

To see whether the protein sample contained antigen, it was resolved by analytical SDS-PAGE, transblotted onto a piece of NitroPlus membrane. The membrane was transferred to blocking buffer (5% skimmed milk powder, 0.2% Tween-20™, 0.02% NaN_3 in 1 x PBS) face up and equilibrated for at least 1 hr at room temperature or overnight at 4 °C. The membrane was then sealed into a plastic bag with 5 ml of blocking buffer containing the first antibody (rabbit anti-antigen) and incubated at room temperature for 1.5 hrs. The membrane was then washed with washing buffer (1 x PBS, 0.5% Tween-20™) three times, each time for 5 min at room temperature. The membrane was sealed into another bag with 5 ml of blocking buffer containing the second antibody (peroxidase-conjugated goat anti-rabbit IgG, BRL) and incubated at room temperature for 1 hr. The membranes was then washed as for the first antibody, then with TBS (0.9% NaCl, 10 mM Tris-HCl, pH 7.5) once at room temperature. The membrane was laid onto a piece of Saran Wrap with equal volumes of LumiGlo Chemiluminescent substrates A and B (KPL) for 1 min under constant mixing. The membrane was finally wrapped in Saran Wrap and exposed to Hyperfilm-ECL film in the dark room for a few sec to 30 min for appropriate signal visualization.

Chapter III. Ribosomal protein operons in the plastid genome of *Guillardia theta*

INTRODUCTION

The secondary endosymbiosis hypothesis is an extension of the endosymbiotic theory for the origin of organelles. A primary endosymbiosis between a photosynthetic prokaryote anJ a phagotrophic eukaryote produced a double-membrane-bound endosymbiont that subsequently became a chloroplast. The resultant chloroplast-containing eukaryote, very similar in structure to red or green algal unicells, was then itself engulfed by another phagotrophic eukaryote. Rather than digesting the engulfed algal cell, the second phagotrophic eukaryote retained the photosynthetic guest, establishing a secondary endosymbiosis. A plastid obtained through secondary endosymbiosis would ha e four membranes, two from the original prokaryotic endosymbiont, a third from the plasma membrane of the primary host, and a fourth from the food vacuolar membrane of the secondary host. The nucleus of the primary host, if it still existed, would lie between the second and third membranes (Figure 3-1; McFadden, 1990).

Cryptomonads are common marine or freshwater biflagellates, with their plastids surrounded by four membranes. Between the inner and outer pairs of membranes surrounding the plastids in cryptomonad cells is a small compartment termed the periplastidal space, which contains ribosome-like particles and a small organelle resembling a nucleus, termed the nucleomorph (Greenwood, 1974). The study of 18S rRNA genes in this organism (Douglas et al., 1991) demonstrated that two species of 18S rRNA genes differing in size by approximately 200 bp are present in this organism. Phylogenetic studies using the 18S rRNA gene sequences revealed that one of them

Figure 3-1. Schematic illustration of secondary endosymbiosis in *Guillardia theta.* A photosynthetic prokaryote (Pr) was engulfed by an unknown eukaryote (Eu1) and became a chloroplast in an unidentified photosynthetic eukaryote. The photosynthetic eukaryote was in turn engulfed by a second phagotroph (Eu2) and became an endosymbiont to create autotrophs such as cryptomonads. Adapted from McFadden (1990).

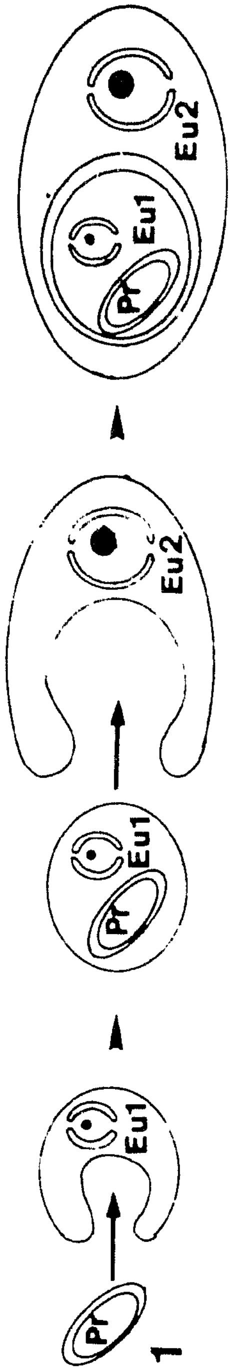


Figure 3-1

clustered with the 18S rRNA genes of rhodophytes, whereas the other was located in a branch containing chlorophytes (land plants, *Chlamydomonas*, etc.) and *Acanthamoeba castellanii*. The existence of two phylogenetically distinct species of 18S rRNAs and the four membranes surrounding the chloroplast clearly demonstrated that *Guillardia theta* obtained its plastid through secondary endosymbiosis.

The endosymbiont hypothesis predicts that organelles were once free living eubacteria. A massive loss of genetic materials and transfer of gen to the nucleus subsequent to endosymbiosis have contributed to the relatively small genome size of the organelles we see today. This genome reduction is also accompanied by the rearrangement of genes remaining in the organelle genome. A major feature of organelle gene organization, in support of this hypothesis, is the presence of eubacteria-like gene operons, such as ribosomal protein operons. Studies of ribosomal protein genes and their organization in the organellar genomes of land plants have shown that remnants of ribosomal protein operons remain in the plastid genomes, although a majority of the genes have presumably been transferred to the nucleus. Only one third of the approximately 60 ribosomal proteins required for plastid ribosomes are encoded by the plastid genomes of land plants (Shinozaki et al., 1986, Ohyama et al., 1986; Hiratsuka et al., 1989) Much less has been known about the organization of plastid ribosomal protein operons in chromophyte algae. Since cryptomonads apparently obtained their plastids through secondary endosymbiosis, it is particularly interesting to see how ribosomal protein genes are organized in their plastid genomes. Data obtained from the studies of ribosomal protein operons should also be valuable in understanding the evolution of the algae and the distribution of ribosomal protein genes between organelle and nuclear genomes.

For this thesis, I sequenced two DNA fragments from the plastid genome of *Guillardia theta*, totaling about 8.3 kbp in size This revealed a total of 17 ribosomal protein genes. Twelve of them encode proteins of the 50S large subunit of ribosomes, five of them encode proteins of the 30S small subunit of ribosomes. A portion of the *secY* gene, which encodes a protein involved in protein targeting and translocation within the cell, was also found at the end of the string of ribosomal protein genes. The arrangement of these genes is *L3-L4-L23-L2-S19-L22-S3-L16-L29-S17-L14-L24-L5-S8-L6-L18-S5-SecY*. This arrangement is identical to corresponding gene arrangement in eubacterial *S10-spc* ribosomal protein operons, except that *rps10, rps14, rpl30* and *rpl15* genes are absent. No introns are present in any of the genes. This is so far the most complete eubacteria-like ribosomal protein operon found in a plastid genome. The plastid genome of *Guillardia theta phi* is highly AT-rich, with 67% A+T and 33% G+C. The codon usage is highly biased toward A and T at the third position. While most of the genes use ATG as their translation initiation codons, the *S3* and *S8* genes use GTG as the start codon. These genes are tightly packed, with intergenic spacers of 2 and 72 bp. The intergenic spacer between *rps17* and *rpl14* is only 5 bp, whereas in *E. coli*, it is 163 bp and separates these two genes into two operons. Northern hybridization analyses detected a 10 kb mRNA, suggesting that these genes are co-transcribed. The evolutionary significance of this unique gene organization is discussed.

RESULTS

A. DNA sequence determination and identification of ribosomal protein genes

The plastid genome of *Guillardia theta* is 118 kbp in size, and its physical map (BamH I and Sal I only) is shown in Fig. 3-2. The region of the genome studied here

Figure 3-2: Cloning of the XS and SS DNA fragments from the plastid genome of *Guillardia theta*. Shown above is the physical map of the plastid genome of *Guillardia theta*. Only the fragments generated with BamH I and Sal I are shown. Restriction endonucleases used in subcloning are shown Numbers in parentheses are relative positions of the sites. Insert in clone BS7 is the 12 kbp BamH I-Sal I fragment, insert in clone S7 is the 2 kbp Sal I-Sal I fragment immediately next to the insert of BS7. XS is an Xba I-Sal I DNA fragment isolated from the 12-kbp BamH I-Sal I DNA fragment, while SS is a 2-kbp Sal I-Sal I DNA fragment located immediately downstream of the XS fragment. The XS and SS DNA fragments were cloned into plasmid vector pUC119.

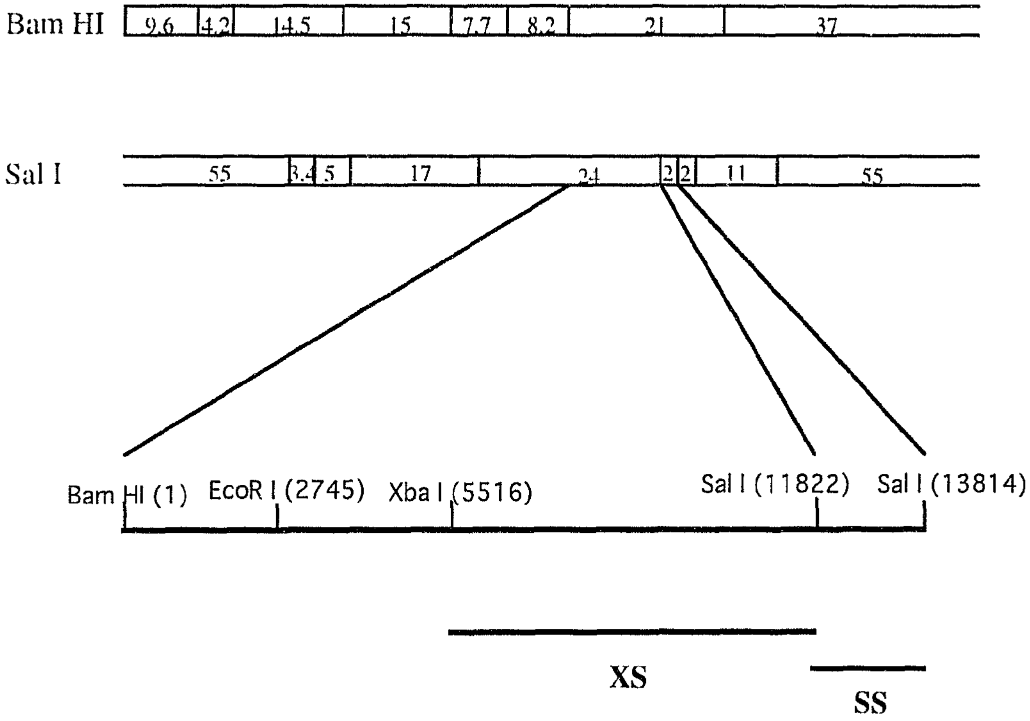

Bam HI
9.6
4.2
14.5
15
7.7
8.2
21
37
Sal I
55
3.4
5
17
24
2
2
11
55
Bam HI (1)
EcoR I (2745)
Xba I (5516)
Sal I (11822)
Sal I (13814)
XS
SS

Figure 3-2

includes the 6.3-kbp Xba I-Sal I DNA fragment (XS) and the adjacent 2-kbp Sal I-Sal I DNA fragment (SS) (Fig. 3-2). Both DNA fragments were cloned in plasmid vector pUC119, and their complete DNA sequences were determined on both DNA strands by the standard dideoxy termination methods.

Complete DNA sequences (Fig. 3-3) of the XS and SS DNA fragments revealed a nucleotide composition of 67% A+T and 33% G+C, indicating that *Guillardia theta* plastid genome is highly AT-rich. This is quite typical of a plastid genome. Sequence analysis revealed 18 open reading frames in this DNA, all of which are encoded on the same DNA strand and transcribed in the direction of Xba I to Sal I. Subsequent analysis of protein sequences deduced from these open reading frames identified them as ribosomal protein genes, since they share significant sequence similarities with known ribosomal proteins of eubacteria and other plastids (see below). The last and incomplete open reading frame is part of the secY gene, which was later described by Douglas (1992). The arrangement of these genes is *L3-L4-L23-L2-S19-L22-S3-L16-L29-S17-L14-L24-L5-S8-L6-L18-S5-SecY*. The *rps3* and *rps8* genes use GTG rather than ATG as their translation initiation codons. Among the translation termination codons used by these 18 genes, 14 are TAA, two are TAG, and two are TGA. There is a strong bias in codon usages. The third codon position is strongly biased toward A and T, as only 16.7% of the codons have G or C at their third position (500 codons with G/C at the third position, including start codon ATGs and GTGs, versus 2491 codons with A/T at the third position). No recognizable introns or stable stem-loop structures was identified. The gene organization is highly compact, with the intergenic spacers ranging from 2 bp to 72 bp.

Amino acid sequence deduced from the coding sequence of each gene was compared with known sequences of homologous proteins from eubacteria, archaebacteria, and other plastids (Fig. 3-4). The sequence alignments clearly identified

Figure 3-3. DNA sequence and deduced protein sequences of ribosomal protein genes from the *Guillardia theta* chloroplast genome. DNA sequences in upper case letters are coding sequences, lower case letters are intergenic sequences. Deduced amino acid sequences are shown beneath the DNA sequences, with the name of each deduced protein shown above the start codon. Restriction sites Xba I and Sal I are shown, with their recognition sites underlined.

```
Xba I        L3
tctagaacattaaattATG AAA ATA GGT TTA TTA GGT ACA AAA TTG GGC ATG ACC CAA ATT TTT GAT
                Met lys ile gly leu leu gly thr lys leu gly met thr gln ile phe asp

GAT AAT GGT TCT GCT ATT CCA GTT ACA ATA TTA AAA GTA GGT CCT TGT TAT GTT ACA AAT CTA
asp asn gly ser ala ile pro val thr ile leu lys val gly pro cys tyr val thr asn leu

AAA TCT GAT ACA AAG GAT AAC TAC AAT GCT ATT CAA ATT GGG TAT CAA CAA GTT GAT GCT AAA
lys ser asp thr lys asp asn tyr asn ala ile gln ile gly tyr gln gln val asp ala lys

AAG TTA ACA AAA CCA CAA TTG GGT CAC TTA CAA GTT AAT AAT TTA CCA CCA TTA AAG CAT TTA
lys leu thr lys pro gln leu gly his leu gln val asn asn leu pro pro leu lys his leu

AAA GAA TAC AAA GTT GAT GCT ACG CAT ACT TTC ACG ATT GCA CAA CAG TTA GAT GTG TCC ATC
lys glu tyr lys val asp ala thr his thr phe thr ile ala gln gln leu asp val ser ile

TTT GAA TTA GGA CAA ATT GTT TCT GTT TCT GGA GTT TCA ATT GGT AAG GGA TTT GCT GGT ACT
phe glu leu gly gln ile val ser val ser gly val ser ile gly lys gly phe ala gly thr

GTA AAG CGA CAT AAT TTT ACA CGT GGT CCG ATG ACA CAT GGA TCG AAA AAT CAT CGT CAA CCA
val lys arg his asn phe thr arg gly pro met thr his gly ser lys asn his arg glu pro

GGT TCG ATT GGA CAA GGT AGT ACA CCG GCT AAA GTT CAT AAA GGT AAA AAA ATG GCT GGT AGA
gly ser ile gly gln gly ser thr pro ala lys val his lys gly lys lys met ala gly arg

TTA GGT GGA CAT CAA GTC ACT ACA AAA AAT TTA ACA GTC GTA CAT TTG GAC AAA GAT AAT AAT
leu gly gly his gln val thr thr lys asn leu thr val val his leu asp lys asp asn asn

GTA TTG GTA CTA AAA GGA TGC GTA CCA GGC AAA CGA GGA AAT ATT CTT AGT ATT AAA TAAaagc
val leu val leu lys gly cys val pro gly lys arg gly asn ile leu ser ile lys OCH
                    L4
tcattaaatacaacatatc ATG GGA TCA AAC AAA ATA AAA AAT TTA GTT CAG TAC GAA ATC CAG GAT
                    Met gly ser asn lys ile lys asn leu val gln tyr glu ile gln asp

TTT GTT TCA TTA AAT AAA ATG GAT ACA AAA TCA CAT GAC TCA CTT AAT TTA AAT GTA AGT AAA
phe val ser leu asn lys met asp thr lys ser his asp ser leu asn leu asn val ser lys

AAA TCA AGA TAT TTA TTA CAT CGC GTT TTA ACA AAT CAA TTA ATT AAT AAT CGA AGT GGT AAT
lys ser arg tyr leu leu his arg val leu thr asn gln leu ile asn asn arg ser gly asn

GCG TGC ACA AAA ACA CGA AGT GAA GTT GAA GGT GGT GGT AAA AAA CCT TGG AAA CAA AAA GGT
ala cys thr lys thr arg ser glu val glu gly gly gly lys lys pro trp lys gln lys gly

ACA GGA AAT GCT AGA GCC GGT TCG AGC AAT TCT CCA CTT TGG AAA GGA GGG GGC GTA ACT TTT
thr gly asn ala arg ala gly ser ser asn ser pro leu trp lys gly gly gly val thr phe

GGT CCT AAA CCT AGA ACC TTT TCA AAT AAG ACT AAT AAA AAG GAA CGA CTT TTA GCT TTA ACA
gly pro lys pro arg thr phe ser asn lys thr asn lys lys glu arg leu leu ala leu thr

ACA GCT TTA TAT TTA AAA TCT AAT AAT ACT AAA GTA ATT AAC TTA GAT AAT TTA GAT TTT ACA
thr ala leu tyr leu lys ser asn asn thr lys val ile asn leu asp asn leu asp phe thr

AAT TTG AAA ACT AGA GAT CTA GTT ATT AAA TGC TCA AAC TTA ATT GAG AAT TAC AAA AAA GAT
asn leu lys thr arg asp leu val ile lys cys ser asn leu ile glu asn tyr lys lys asp

CAA AAA ATT CTT TTT GTT GCT GAA CCT ACT GCC AGT GGT CTT TGG AGA TAT GTA AAA AAT ATA
gln lys ile leu phe val ala glu pro thr ala ser gly leu trp arg tyr val lys asn ile
```

Figure 3-3 continues

```
TCT AAT GTT GAT TTA ATT TAC ACG ACT GGT CTA GAT CTG AAA AAA ATT CTA CAA GCA CAT CAT
ser asn val asp leu ile tyr thr thr gly leu asp leu lys lys ile leu gln ala his his

ATA ATT TTT ACG TGT AAA GCA CTA AAT GAT GTT AAG GAG GTA TTC AAT GAA CAA TAA acacaaa
ile ile phe thr cys lys ala leu asn asp val lys glu val phe asn glu gln OCH
    L23
gat ATG CAT GCT TTA ATT GAT TTA GTA AAG TAT CCA TTA ATT ACT GAT AAA GCC ACA AGA TTA
    Met his ala leu ile asp leu val lys tyr pro leu ile thr asp lys ala thr arg leu

CTT GAG TTA AAT CAA TAC ACT TTT TTA ACT TCT CGT GTT GCT ACA AAA ACA GAT ATA AAA AAT
leu glu leu asn gln tyr thr phe leu thr ser arg val ala thr lys thr asp ile lys asn

GCT ATT GAA TTT TTA TTT AAT GTA AAA GTA ATA AGT ATC AAT ACA TGT TTG TTA CCA TTA AAA
ala ile glu phe leu phe asn val lys val ile ser ile asn thr cys leu leu pro leu lys

CGT AAA AGA TTA GGT AAG TTC GTA GGC TCA AAA CCT CGT TAT AAA AAA GCT GTT GTT ACG TTA
arg lys arg leu gly lys phe val gly ser lys pro arg tyr lys lys ala val val thr leu
                                                                    L2
GAA AAA AAT AAT ACA ATT AAC CTA TTT TCT GAA AAT TAA ataaattaacatttt ATG GGA ATA CGC
glu lys asn asn thr ile asn leu phe ser glu asn OCH                 Met gly ile arg

ATC TAT AAA TCT TAT ACT CCA GGT ACT CGA AAT CGA TCT AGT TCT GAC TTT GTT GAA ATT ACA
ile tyr lys ser tyr thr pro gly thr arg asn arg ser ser ser asp phe val glu ile thr

AAA TCA AAA CCA GAA AAA TCA TTA CTC CGT AAA AAA TTG TCT TGT GCA GGT AGA AAT AAT CGT
lys ser lys pro glu lys ser leu leu arg lys lys leu ser cys ala gly arg asn asn arg

GGT TTA ATA ACC GTA CGG CAC AAA GGA GGT GGA CAT AAA CAA CGT TAT CGA TTG GTC GAT TTT
gly leu ile thr val arg his lys gly gly gly his lys gln arg tyr arg leu val asp phe

AAA CGT AAT AAA TTG GAT ATA CCT GCT ATT GTC GCA TCT GTC GAA TAT GAT CCA AAT CGA AAT
lys arg asn lys leu asp ile pro ala ile val ala ser val glu tyr asp pro asn arg asn

GCC AGA ATT GCC CTA CTA CAT TAT CAA GAT GGT GAA AAA CGT TAT ATC TTA CAT CCT AAA AAA
ala arg ile ala leu leu his tyr gln asp gly glu lys arg tyr ile leu his pro lys lys

TTG GCA GTG GGA GAT AAA ATA TAT TCA GGT ATT AAT GTA CCT ATA GAA ATT GGT AAT GCA ATG
leu ala val gly asp lys ile tyr ser gly ile asn val pro ile glu ile gly asn ala met

CCA TTA TAT AAT GTC CCA TTA GGT ACT GCT GTT CAC AAT GTT GAA CTA ATA CCG GGA CGA GGT
pro leu tyr asn val pro leu gly thr ala val his asn val glu leu ile pro gly arg gly

GGT CAA ATT GTG CGC TCA GCA GGA ACT TCT GCA CAA GTC GTA GCA AAA GAT GGA CAA GTT GTA
gly gln ile val arg ser ala gly thr ser ala gln val val ala lys asp gly gln val val

ACT ATA AAG ATG CCA TCT AAT GAA GTA CGC ATG ATT TAT AAA AAT TGT TAT GCA ACT ATT GGT
thr ile lys met pro ser asn glu val arg met ile tyr lys asn cys tyr ala thr ile gly

GAA GTA GGA AAT GCA GAT ATT AAA AAT ATT CGT TTA GGT AAA GCG GGA CGA AAA CGG TGG TTG
glu val gly asn ala asp ile lys asn ile arg leu gly lys ala gly arg lys arg trp leu

GGG ATT CGT CCA TCT GTT AGA GGT GTA GTA ATG AAT CCT TGT GAT CAC CCT CAT GGT GGT GGT
gly ile arg pro ser val arg gly val val met asn pro cys asp his pro his gly gly gly

GAA GGT CGT TCT CCC ATT GGT AGA GCA AAG CCA GTT ACT CCT TGG GGT AAG CCT GCT TTA GGT
glu gly arg ser pro ile gly arg ala lys pro val thr pro trp gly lys pro ala leu gly
```

Figure 3-3 continues

```
GTA AAG ACA CGG AGA CAG AAT AAA TAT AGT GAT TTT TGT ATA ATA CGA TCA CGT AAT TAAacta
val lys thr arg arg gln asn lys tyr ser asp phe cys ile ile arg ser arg asn OCH
                        S19
taaccttaaataatatattcaaatATG AGT AGA TCT TTA TCT AAA GGC CCA TAT ATT GCG GCT CAT TTA
                        Met ser arg ser leu ser lys gly pro tyr ile ala ala his leu

TTA AAA AAG TTG AAT AAT GTT GAT ATT CAA AAA CCT GAT GTT GTT ATA AAA ACT TGG TCT CGT
leu lys lys leu asn asn val asp ile gln lys pro asp val val ile lys thr trp ser arg

TCA TCA ACC ATA TTA CCT AAC ATG GTT GGA GCG ACA ATT GCT GTT TAT AAC GGT AAA CAA CAT
ser ser thr ile leu pro asn met val gly ala thr ile ala val tyr asn gly lys gln his

GTG CCC GTT TAT ATT TCG GAT CAA ATG GTT GGA CAC AAA TTA GGG GAA TTT TCG CCT ACT CGT
val pro val tyr ile ser asp gln met val gly his lys leu gly glu phe ser pro thr arg

ACA TTT AGG TCC CAT ATC AAA AGT GAT AAA AAA GCA AAA CGT TAA tttataaatttatttcgattcca
thr phe arg ser his ile lys ser asp lys lys ala lys arg OCH
    L22
aat ATG ATA CTA TCA CTA AAT TCA CCC AAT GTC GCC GTG CCA ACA GCG AAA TAT ATT CGA ATG
    Met ile leu ser leu asn ser pro asn val ala val pro thr ala lys tyr ile arg met

TCG CCC TCA AAA ATA CAA CGT GTT TTA AAT CAA ATT CGT GGT AAA TCC TAT AAA GAA AGT CTA
ser pro ser lys ile gln arg val leu asn gln ile arg gly lys ser tyr lys glu ser leu

ATG ATA CTA GAA TTT ATG CCT TAT GCA GCT TGC AAA CCT GTA TTG CAA GCT GTT CAG TCA GCA
met ile leu glu phe met pro tyr ala ala cys lys pro val leu gln ala val gln ser ala

GGA GCT AAT GCT CAA CAT AAT AAA GGG ATT AAT AAA AAT GAT TTA GTT GTT TCT TTA GCT TCT
gly ala asn ala gln his asn lys gly ile asn lys asn asp leu val val ser leu ala ser

GTG GAT AAT GGA CCA GTT CTT CGC CGT TTT AGA CCT AGA GCC CAA GGA AGA GGA TTT AAA ATA
val asp asn gly pro val leu arg arg phe arg pro arg ala gln gly arg gly phe lys ile

CAA AAG TTT ACT TCG CAT ATA CGA ATC GGA GTA CAA AAA CAA GTT AAT TTT TAA taaacatgtaa
gln lys phe thr ser his ile arg ile gly val gln lys gln val asn phe OCH
                    S3
ataaaaaaaggagttatt  GTG GGA CAA AAA GTA AAT CCA TTA GGT TTT AGA CTT CGT ATT ACA AGT
                    Met gly gln lys val asn pro leu gly phe arg leu arg ile thr ser

CAA CAT CGT TCA TCA TGG TTC GCA ACT AAA GAA TCG TAT CCT CAA TTG TTA GAA CAG GAT TTT
gln his arg ser ser trp phe ala thr lys glu ser tyr pro gln leu leu glu gln asp phe

AAG ATT CGA TCA TAT ATT AAT CGT GAA TTG GAA GCT GCT GGA ATT TCA AAA ATT GAA ATA AGT
lys ile arg ser tyr ile asn arg glu leu glu ala ala gly ile ser lys ile glu ile ser

CGA AAT GCA AAT CAA TTA GAA GTG TCC GTT TAT ACT TCA AGA CCC GGT ATA ATT GTA GGT CGT
arg asn ala asn gln leu glu val ser val tyr thr ser arg pro gly ile ile val gly arg

TCT GGT CTT GGT ATC GAA AAA ATA AAA ACA GAT ATA CTA CGT TTA TTA AAA CAA GAT ATA TCA
ser gly leu gly ile glu lys ile lys thr asp ile leu arg leu leu lys gln asp ile ser

ATC CGG ATT AAT GTT ATA GAA TTA ACT AAT CCA GAT GCT GAT GCA AAC TTA ATT GGT GAA TTT
ile arg ile asn val ile glu leu thr asn pro asp ala asp ala asn leu ile gly glu phe

ATT GCT CAA CAA CTA GAA AAG CGC GTT GCG TTT CGT CGG GCT ACA CGA CAA GCA ATT CAA AAA
ile ala gln gln leu glu lys arg val ala phe arg arg ala thr arg gln ala ile gln lys
```

Figure 3-3 continues

```
GCA CAA CGA GCT AAC GTA CAA GGT ATA AAA GTT CAA GTA TCA GGC CGA TTA AAT GGA GCT GAA
ala gln arg ala asn val gln gly ile lys val gln val ser gly arg leu asn gly ala glu

ATA GCA CGT AGT GAA TGG GTC CGT GAA GGT AGA GTT CCA CTG CAA ACT TTA AGA GCA AAT ATA
ile ala arg ser glu trp val arg glu gly arg val pro leu gln thr leu arg ala asn ile

GAT TAT GCT ACT AAA GAA GCT CAC ACA ACC TAT GGT ATC CTG GGT ATA AAA GTT TGG GTA TTT
asp tyr ala thr lys glu ala his thr thr tyr gly ile leu gly ile lys val trp val phe

AAC GGT GAA CAG ACA CCA ACG TAT GCT GTC ATT TAGatgaaatggcaaatttagtcaataaaaaaatataaa
asn gly glu gln thr pro thr tyr ala val ile AMB
                L16
cgatattatcaaatcaATG CTT AGT CCT AAA AGA ACA AAG TTT CGT AAA CCG CAT AGA GGA AGA TTA
                Met leu ser pro lys arg thr lys phe arg lys pro his arg gly arg leu

CGT GGT ATA GCT ACT AGA GGA AAC ACA CTT ATA TTT GGC GAT TAT GGT CTT CAA GCT TTA GAA
arg gly ile ala thr arg gly asn thr leu ile phe gly asp tyr gly leu gln ala leu glu

CCT ATA TGG TTA ACT TCA AGA CAA ATT GAA GCT ACA CGA CGC ACA ATA ACA AGA CAA GTT AAA
pro ile trp leu thr ser arg gln ile glu ala thr arg arg thr ile thr arg gln val lys

CGA GTT GGT CGA TTG TGG ATT CGT GTA TTT CCT GAT AAA TCT ATT TCT GCT AAA CCA CCA GAA
arg val gly arg leu trp ile arg val phe pro asp lys ser ile ser ala lys pro pro glu

ACA AGA ATG GGA GCT GGA AAA GGT GCC CCA GAA TAT TGG GTA GCT GTA ATT AAA CCT GGT CAT
thr arg met gly ala gly lys gly ala pro glu tyr trp val ala val ile lys pro gly his

ATA TTG TTT GAA ATT AAT GGT GTA TCA CAG GAT TTA CGT TAT TTG GCT TTT AAA AAT GCT TCT
ile leu phe glu ile asn gly val ser gln asp leu arg tyr leu ala phe lys asn ala ser
                                                                L29
TAC AAA TTA CCT ATA AAA ACA AAA TTT ATT TCA CGT TAA acaatatttt  ATG ACT ACA AAT TTA
tyr lys leu pro ile lys thr lys phe ile ser arg OCH             Met thr thr asn leu

GAT TCT ACA CAG TTA GAA AAA TTA ACA GAT ACT GAT ATT AAT GAT ACA GTG TTG AAA TTA AAA
asp ser thr gln leu glu lys leu thr asp thr asp ile asn asp thr val leu lys leu lys

AAA GAA TTA TTC GAA TTG AGA CTT CAA AAA GCA ACA AGA CAG GAA ATA AAA CCA CAT TTA TTT
lys glu leu phe glu leu arg leu gln lys ala thr arg gln glu ile lys pro his leu phe

AAG CAA AAG AAA AAA TTA ATT GCT AAA CTT CTA ACA ATA AAA TCA AAA AAA AGT TAA atattta
lys gln lys lys lys leu ile ala lys leu leu thr ile lys ser lys lys ser OCH
                S17
aattaggagaatatgtATG TCT ATT AAA GAG AGG TTA GGG TTA GTC ATT AGT GAC AAA ATG GAT AAA
                Met ser ile lys glu arg leu gly leu val ile ser asp lys met asp lys

ACC GTT GTT GTT TCG ATT GCA AAT CGC GTT ACT CAT AAA CGT TAC GGG AAG ATT GTT ACA AAA
thr val val val ser ile ala asn arg val thr his lys arg tyr gly lys ile val thr lys

ACA AAA AAG TAT AAG GTT CAT GAT CCT AAT AAT AAT TGT CAA GTG GGT GAT TTA ATT TTG ATA
thr lys lys tyr lys val his asp pro asn asn asn cys gln val gly asp leu ile leu ile

AAT GAA ACA CGT CCA TTG AGT AAG ACT AAA CGT TGG ATG TTT AAA GAA ATC AAA CAA AAA TCT
asn glu thr arg pro leu ser lys thr lys arg trp met phe lys glu ile lys gln lys ser
                                                L14
TTA AAA TTG GAT AAA GAT ACA ATT GGA GAA TAAattatATG ATT CAA ACT CAG ACT TAT TTA ACT
leu lys leu asp lys asp thr ile gly glu OCH     Met ile gln thr gln thr tyr leu thr
```

Figure 3-3 continues

```
GTA GCT GAT AAC AGT GGA GCT AAA AAA ATA ATG TGT ATA CGG ATT TTG GGA GGT AAT AGA AAA
val ala asp asn ser gly ala lys lys ile met cys ile arg ile leu gly gly asn arg lys

TAT GCT TCC ATA GGT GAT GTA ATT ATC GGT GTT GTA AAA GAT GCT ACT CCA AAC ATG CCG GTA
tyr ala ser ile gly asp val ile ile gly val val lys asp ala thr pro asn met pro val

AAA CGA TCA GAT GTT GTA CGT GCA GTT ATT ATG AGA ACA AAA AAT ACT ATA CGA CGT AAA GAT
lys arg ser asp val val arg ala val ile met arg thr lys asn thr ile arg arg lys asp

GCA ATG TCT ATA AGA TTT GAT GAT AAC GCT GCA GTT ATT ATA AAT AAA GAA AAT AAT CCA CGT
gly met ser ile arg phe asp asp asn ala ala val ile ile asn lys glu asn asn pro arg

GGT ACA AGA GTA TTT GGC CCT ATT GCT AGA GAA CTA CGT GAT AAA GAT TTT ACA AAA ATT GTT
gly thr arg val phe gly pro ile ala arg glu leu arg asp lys asp phe thr lys ile val
                                    L24
TCC TTA GCT CCA GAG GTA TTA TGA aa  ATG ACA ATA AAA CAA GGT GAT AAA GTT CAA GTT ATT
ser leu ala pro glu val leu OPA     Met thr ile lys gln gly asp lys val gln val ile

GCA GGA AGT TAT AAA GGT GAA ATT ACG GAA GTT TTA AAA GTA ATT CGT AAA TCT AAT TCT TTA
ala gly ser tyr lys gly glu ile thr glu val leu lys val ile arg lys ser asn ser leu

ATT TTG AAA AAT ATT AAT ATA AAA AAT AAA CAT GTT AAG CCA AAA AAA GAA GGT GAA GTA GGT
ile leu lys asn ile asn ile lys asn lys his val lys pro lys lys glu gly glu val gly

CAA ATT AAA CAG TTT GAA GCT CCT ATT CAT CGA TCA AAT GTT ATG TTA TAT GAT GAA GAA TCA
gln ile lys gln phe glu ala pro ile his arg ser asn val met leu tyr asp glu glu ser

CAA ATT CGT AGT CGT AGT AAA TTT ATA ATT AGC CAA GAT GGC AAA AAA GTT AGA GTT TTA AAG
gln ile arg ser arg ser lys phe ile ile ser gln asp gly lys lys val arg val leu lys
                                L5
AAA TTA GTA AAA AAT TAGacatgtactATG AAT AAA AGT TTA AAA GAA ATA TAT TAT CAA GAC GTT
lys leu val lys asn AMB         Met asn lys ser leu lys glu ile tyr tyr gln asp val

ATA CCG GGT TTA ATT GAA CAA TTT AAT TAT ACT AAT ATT CAC CAA GTT CTT AAA ATT ACA AAA
ile pro gly leu ile glu gln phe asn tyr thr asn ile his gln val leu lys ile thr lys

ATA ACT TTA AAT CGT GGT CTT GGC GAA GCT TCA AAA AAT AAC AAA ATT TTA GAA GCT AGT ATT
ile thr leu asn arg gly leu gly glu ala ser lys asn asn lys ile leu glu ala ser ile

AAG GAG TTT GAA CTA ATT TCT GGT CAA CAC CCA CTA ATA AAT AAA GCT CGT AAA TCT GTT GCC
lys glu phe glu leu ile ser gly gln his pro leu ile asn lys ala arg lys ser val ala

GGA TTT AAA ATT AGA GAA GGT ATG CCT GTT GGT ATA TCA GTA ACC TTA CGT AAA AAA TTG ATG
gly phe lys ile arg glu gly met pro val gly ile ser val thr leu arg lys lys leu met

TAT ACA TTT TTA GAA AAA CTG ATT CAT CTT TCT TTA CCG CGT ATT CGT GAT TTT AGA GGA GTT
tyr thr phe leu glu lys leu ile his leu ser leu pro arg ile arg asp phe arg gly val

AGT GTA AAA AGT TTT GAT GGT CGA GGA AAT TAT AAT TTA GGT ATT AAA GAG CAA TTA ATC TTT
ser val lys ser phe asp gly arg gly asn tyr asn leu gly ile lys glu gln leu ile phe

CCA GAA ATT GAA TAT GAT CAA GTT GAT CAG GTT CGT GGT TTA GAT ATC TCT ATA ACA ACC ACC
pro glu ile glu tyr asp gln val asp gln val arg gly leu asp ile ser ile thr thr thr

GCT AAA ACA CAA CAA GAA GGA ATT GCC CTT CTG CGA GCA TTA GGC ATG CCA TTT AAT GAT AAT
ala lys thr gln gln glu gly ile ala leu leu arg ala leu gly met pro phe asn asp asn
```

Figure 3-3 continues

```
                                        S8
TAA ttt taa ttt aaa ttc aaa gag gta att GTG ACA AAT GAT ACT GTC TCA GAC ATG [illegible] ACA
OCH                                     Met thr asn asp thr val ser asp met leu thr

AGA GTT CGA AAT GCA AAT TTA GCT AAA CAC CAA GTT GTG CAG GTG CCA GCA ACA AAA ATG ACA
arg val arg asn ala asn leu ala lys his gln val val gln val pro ala thr lys met thr

AAA AGT ATT GCA CAC GTA TTA CTA GAA GAA GGG TTC ATC GAA AGT ATT GAA GAA GTT GGT TTA
lys ser ile ala his val leu leu glu glu gly phe ile glu ser ile glu glu val gly leu

GAT ATA AAT AGA CAA TTA TTA CTT TCT CTT AAG TAT AAA GGA AGG GAA AGA GAA CCG GTT ATA
asp ile asn arg gln leu leu leu ser leu lys tyr lys gly arg glu arg glu pro val ile
                        Sal I
AAT GCG TTA AAA AGA ATT AGT CGA CCT GGC TTA CGA GTA TAT GCG AAT CGT AAA GAG TTA CCA
asn ala leu lys arg ile ser arg pro gly leu arg val tyr ala asn arg lys glu leu pro

CGT GTT TTA GGT GGT TTA GGA ATC GCT GTG ATT TCA ACA TCG AAG GGA GTC TTA ACT GAT ACA
arg val leu gly gly leu gly ile ala val ile ser thr ser lys gly val leu thr asp thr

AAA GCT AGA ACA CAA GGT CTT GGT GGT GAA GTT TTA TGT TAC ATC TGG TAA ttaatagaaaatata
lys ala arg thr gln gly leu gly gly glu val leu cys tyr ile trp OCH
                        L6
agttaacttaagaggtaaaaac  ATG TCA AGA ATT GGA AAG CTT CCT GTT AAA TTT TCA GAG AAA GTT
                        Met ser arg ile gly lys leu pro val lys phe ser glu lys val

ACT ATG AAA ATT GAT CAG GAC AAT ATT ATT GTA AAA GGT CCA AAA GGA GAA TTA GCG TTG GGT
thr met lys ile asp gln asp asn ile ile val lys gly pro lys gly glu leu ala leu gly

CTT TCA AAA AAT ATT AAT ATT ACA ATT GAA AAT AAT ACG TTG TTT GTA AAA CCT GTT ACA AAA
leu ser lys asn ile asn ile thr ile glu asn asn thr leu phe val lys pro val thr lys

GAA CCC CAA GTC CTT AAA TTA TTT GGA ACG TAC AGA GCG ATC ATT AAT AAT ATG GTT GTT GGG
glu pro gln val leu lys leu phe gly thr tyr arg ala ile ile asn asn met val val gly

GTA ACT AAA GGT TTT GAG AAG AGA CTT GAA CTA CAA GGG GTG GGC TAT CGT GCA CAG CTT CAG
val thr lys gly phe glu lys arg leu glu leu gln gly val gly tyr arg ala gln leu gln

GGA AAA GAT TTG TCT TTA AGT GTT GGT TAT AGT CAT CCA GTT GTA ATA AAA GCT CCG ACT GGA
gly lys asp leu ser leu ser val gly tyr ser his pro val val ile lys ala pro thr gly

ATC AAT ATT GCG GTT GAA AAT AAT ACT ATT GTT ATT ATA TCT GGT ATT AGT AAG GAA TTA GTT
ile asn ile ala val glu asn asn thr ile val ile ile ser gly ile ser lys glu leu val

GGT CAA ATT GCG TCA AAT ATA AGA TCT ATA AAA CCT CCA GAA CCT TAC AAA GGT AAA GGA ATA
gly gln ile ala ser asn ile arg ser ile lys pro pro glu pro tyr lys gly lys gly ile

AGA TAT GTT GGT GAA TTT GTA CGT AAA AAA GCT GGA AAA GCA GGT AAA AAA TAA ttcaaaatgc
arg tyr val gly glu phe val arg lys lys ala gly lys ala gly lys lys OCH
L18
ATG AAA AGA ACA AAT AAA ATT AAG GGA ACA TTA GAA CGA CCA CGT TTA TCT GTA TTT CGT TCA
Met lys arg thr asn lys ile lys gly thr leu glu arg pro arg leu ser val phe arg ser

AAT TGT CAT ATT TAC GCA CAA GTA ATT GAT GAT TCT TCT GGT ATG ACT ATT GTA TCA ACT TCA
asn cys his ile tyr ala gln val ile asp asp ser ser gly met thr ile val ser thr ser

ACT TTA GAC AAA GAT GTT AAA AGT TTA TTG AAC AAT ACT TCA ACT TGT GAA GCT TCG AAA ATT
thr leu asp lys asp val lys ser leu leu asn asn thr ser thr cys glu ala ser lys ile
```

Figure 3-3 continues

```
GTA GGG CAA GTC ATC GCG AAA AAG ACA CTT GCA CGA AAT ATA AAA CAA GTA ATT TTT GAT AGA
val gly gln val ile ala lys lys thr leu ala arg asn ile lys gln val ile phe asp arg

GGA AAA CGT GTC TAC CAT GGT AGA ATT TCT GCT TTA GCA GAA GCC GCA CGA GAA AGC GGA TTA
gly lys arg val tyr his gly arg ile ser ala leu ala glu ala ala arg glu ser gly leu
                                                S5
GAA TTT TAA att tat aca aaa tat gaa cat ctt att ATG TTA AAC GCA AAA AAA TCG AAT AAA
glu phe OCH                                     Met leu asn ala lys lys ser asn lys

ACT AAA GAG AAA GAA ACT GAT TGG CAA GAG CGT GTT ATT CAA GTG CGA CGT GTT ACT AAG GTT
thr lys glu lys glu thr asp trp gln glu arg val ile gln val arg arg val thr lys val

GTT AAA GGT GGT AAA AAA TTA AGC TTC AGA GCT ATT ATT ATT TTG GGT AAT GAG CGT GGA CAG
val lys gly gly lys lys leu ser phe arg ala ile ile ile leu gly asn glu arg gly gln

GTT GGC GTA GGT GTA GGT AAA GCA AGT GAC GTG ATA GGG GCA GTA AAA AAA GCA GTG ACA GAC
val gly val gly val gly lys ala ser asp val ile gly ala val lys lys ala val thr asp

GGT AGA AAG AAT CTG ATC AAC ATT CCT TTA ACG AAT CAA AAC TCA ATA CCA CAT ATT GTT CAA
gly arg lys asn leu ile asn ile pro leu thr asn gln asn ser ile pro his ile val gln

GGT TAT TCT GGT GCT GCT AAA GTT ATA ATT AAA CCT TCT GCA CCA GGC TCC GGT GTA ATC GCA
gly tyr ser gly ala ala lys val ile ile lys pro ser ala pro gly ser gly val ile ala

GGT GGT TCT GTC AGA ACT ATT TTA GAA TTA GCT GGG ATA AAA AAT ATC TTA GCT AAG CAA CTT
gly gly ser val arg thr ile leu glu leu ala gly ile lys asn ile leu ala lys gln leu

GGT TCG TCA AAT CCA CTA AAT AAC GCT CGT GCA GCA GCT AAT GCG TTG ATT AAT TTA CGC ACT
gly ser ser asn pro leu asn asn ala arg ala ala ala asn ala leu ile asn leu arg thr

TAT ACA AGT GTT CTA AAT GAT CGT AAC TTA GAC CTT CAT TGAtgttttttcgtgttataagatttctaa
tyr thr ser val leu asn asp arg asn leu asp leu his OPA
                                            SecY
attgagattagaaatcttataatcatatatttaccaatacgtttATG AAT ACC TCA ATA AAA TCT ATT AAA AAA
                                            Met asn thr ser ile lys ser ile lys lys

CAA GAT TTA AAA GAT CGG ATA GTA TTT ACG TTG TTT TTA ATT GTC ATG TCT CGT TTA GGT ACA
gln asp leu lys asp arg ile val phe thr leu phe leu ile val met ser arg leu gly thr

TTT CTG CCC ATA CCG GGA GTT GAT CAT GAT GCT TTT TAT CAA AGT ATA ATA AGT AAT CCA TTA
phe leu pro ile pro gly val asp his asp ala phe tyr gln ser ile ile ser asn pro leu

GTT AAT TTT CTA AAT GTA TTT TCT GGA GGT GGG TTT GCT TCG ATC GGT GTT TTT GCT TTA GGT
val asn phe leu asn val phe ser gly gly gly phe ala ser ile gly val phe ala leu gly

ATA GTT CCT TAC ATA AAT GCT TCA ATT ATT GTA CAA TTA GCT ACT AAT TCG ATC CCG AGT TTA
ile val pro tyr ile asn ala ser ile ile val gln leu ala thr asn ser ile pro ser leu
                                          Sal I
GAA AAG TTA CAA AAA GAA GAA GGT GAA TTA GGT CGA CAA AAA ATA GTT CAA CTT ACA AGA TAT
glu lys leu gln lys glu glu gly glu leu gly arg gln lys ile val gln leu thr arg tyr

GTG GCA TTA GTG TGG GCT TTG ATT CAA AGT ATT GGA GTA TCA TTT TGG GTA CGA CCT TAT GTA
val ala leu val trp ala leu ile gln ser ile gly val ser phe trp val arg pro tyr val

TTT AAC TGG GAT TTA AAC TTT GTT TTC GCT ATG AGC TTA ACC TTA ACT ATA GGT TCG ATG TTA
phe asn trp asp leu asn phe val phe ala met ser leu thr leu thr ile gly ser met leu
```

Figure 3-3 continues

```
ATA ATG TGG TTT TCA GAA CAA ATA ACT GAA AAA GGA ATA GGT AAT GGT CCT TCA CTA CTT ATT
ile met trp phe ser glu gln ile thr glu lys gly ile gly asn gly pro ser leu leu ile

TTT ATT AAT ATT ATT TCT GGA TTA CCT AAA TTG TTA CAA TCA CAA ATT CAA TCC ACT CGT CTT
phe ile asn ile ile ser gly leu pro lys leu leu gln ser gln ile gln ser thr arg leu

AAT ATT CAA GCA TTA GAT ATA TTT GTA CTT GTT TTC ATT TTT TCA GTC ATG ATA ATT GGG ATT
asn ile gln ala leu asp ile phe val leu val phe ile phe ser val met ile ile gly ile

ATT TTT ATA CAA GAA GGT ATA AAA CGA ATT CCT ATC ATT TCT GCA CGG CAA CTT GGT AAA GGG
ile phe ile gln glu gly ile lys arg ile pro ile ile ser ala arg gln leu gly lys gly

CAA ATG GAT AAT AAA ACA AGT TAT TTA CCT TTG AAA CTG AAT CAA AGC GGT GTA ATG CCA ATT
gln met asp asn lys thr ser tyr leu pro leu lys leu asn gln ser gly val met pro ile

ATA TTT GCC TCT GCT GTT TTA GTC TTA CCA GCT TAT TTA GCC CAA CTG GTA TCG AAT GAA CAA
ile phe ala ser ala val leu val leu pro ala tyr leu ala gln leu val ser asn glu gln

TTA AGA ACA GTC TTA CAT TTG TTT GAT GGT ACG AGT AAT AAT AAA TTA CTT TAT TTA TTA TTC
leu arg thr val leu his leu phe asp gly thr ser asn asn lys leu leu tyr leu leu phe

TAT TTT ACA TTA ATT TTA TTC TTT AGT TAT TTT TAT ACA TCT TTA ATA TTG AAT CCA AAT GAT
tyr phe thr leu ile leu phe phe ser tyr phe tyr thr ser leu ile leu asn pro asn asp

GTA TCC AAA AAT CTG AAA AAA ATG GAG TCT AGT ATT TAT GGT GTT CGA CCA GGT AAA GCT ACT
val ser lys asn leu lys lys met glu ser ser ile tyr gly val arg pro gly lys ala thr

ACA GAA TAT TTA CAA AAA ACA TTG AAT CGA CTA ACA TTT TTA GGA GCT TTA TTC TTG GCT TTT
thr glu tyr leu gln lys thr leu asn arg leu thr phe leu gly ala leu phe leu ala phe

ATA GCT ATT GTT CCT AAT ATT ATT GAA ACA TTA ACT AAT TTA TCT GTA TTT AAA GGT TTA GGT
ile ala ile val pro asn ile ile glu thr leu thr asn leu ser val phe lys gly leu gly

GGT ACC TCA TTA TTA ATA ATT GTT GGC GTA CAA GTT GAC ACC TCT AAG CAA ATT CAA ACT TAT
gly thr ser leu leu ile ile val gly val gln val asp thr ser lys gln ile gln thr tyr
                                                                            L36
CTG ATT TCA AAA AAT TAT GAA ACT ATA GTA CGT TAA cttaaatttaaataaatatattaattc ATG AAA
leu ile ser lys asn tyr glu thr ile val arg OCH                             Met lys

GTA GTA AGT TCA ATT GGT AGT TTA AAA AAT CGT AGT AAA GAT TGT CAA ATA GTT AAA AGA AGA
val val ser ser ile gly ser leu lys asn arg ser lys asp cys gln ile val lys arg arg

GGT CGA ATT TAC GTT ATT TGT AAA ACT GAT CCA CGA CTT AAA GTT CGC CAA GGT GGA GCA AAA
gly arg ile tyr val ile cys lys thr asp pro arg leu lys val arg gln gly gly ala lys

ATG AAA CGT AAA TAA tgacctttataaattaaattactgtataaaaaattttaggagaaacata
met lys arg lys OCH
```

Figure 3-3

these *Guillardia theta* plastid genes as the designated ribosomal protein genes. The same sequence alignment also revealed that GTG is the initiation codon in the *rps3* and *rps8* genes. This assignment of GTG as the initiation codon for *rps3* and *rps8* gene is further supported by the finding of putative ribosome binding site preceding these genes (Table 3-1). Putative ribosome binding sites were also identified for genes *rpl3*, *rpl29*, *rpl14*, *rpl24*, and *rpl6*. For the remaining genes, a putative ribosome binding site was either found more distantly upstream or not found at all (Table 3-1). Interestingly, similar studies in the cyanelle genome of *Cyanophora paradoxa* also identified GTG as initiation codon of *rps3* and *rps8* genes (Christine et al., 1990), while all the other genes use ATG as start codon.

Based on the amino acid sequence alignments of homologous proteins from *Guillardia theta* and other organisms (cyanelle, liverwort, Euglena, rice, tobacco, and eubacteria), the percentage of sequence identity was calculated and listed in Table 3-2. In general, the *Guillardia theta* plastid ribosomal proteins show higher sequence identities to corresponding proteins from cyanelle (*Cyanophora paradoxa*) than from other organisms.

B. Ribosomal protein gene organization, operon structure and evolution

When the organization of the ribosomal protein gene cluster in *Guillardia theta* was compared to that of ribosomal protein gene operons in *E. coli* and that of homologous gene clusters in other organisms, the *Guillardia theta* plastid genome was found to contain most of the genes found in eubacterial *S10* and *spc* operons (Fig. 3-5). The *S10* and *spc* operons in *E. coli* together encode 23 genes. Out of these 23 genes, 19 were found in the plastid genome of *Guillardia theta*. whereas 15 are encoded in the cyanelle of *Cyanophora paradoxa*, 10 are found in the chloroplast of *Euglena gracilis*,

Figure 3-4. Protein sequen.-e comparisons. Each protein sequence deduced from the *Guillardia theta* plastid DNA sequence is aligned with corresponding ribosomal protein sequences of other organisms. Organisms compared include *Guillardia theta* (CRYPT), *Cyanophora paradoxa* (CYAPA), *Gracilaria tenuistipitata* (GRATE), *Marchantia polymorpha* (MARPO), *Chlamydomonas reinhardtii* (CHLRE), *Chlamydomonas species* (strain WXM) (CHLSP), *Euglena gracilis* (EUGRA), *Zea mays* (MAIZE), *Oryza sativa* (ORYSA), *Nicotiana tabacum* (TOBAC), *Spinacia oleracea* (SPIOL), *Sinapsis alba* (SINAL), *Astasia longa* (ASTLO), *Mycoplasma* (MYCOP), *Yersinia* (YERSI), *Bacillus subtilis* (BACSU), *Bacillus stearothermophilus* (BACST), *Escherichia coli* (ECOLI), *Mycobacterium leprae* (MYCOL), *Halobacterium marismortui* (HALMA), *Thermoplasma* (THERM). The amino acid sequences were aligned by using the sequence alignment program of Clustal V (Higgins et al. 1992).

RPL3

```
RL3_CRYPT   MKIGLLGTKL GMTQIFDDNG SAIPVTILKV GPCYVTNLKS DTKDNYNAIQ
RL3_CYAPA   .S..I..... .......EA. N......IQA ...PI.QI.T TAT.G.....
RL3_ECOLI    M...V.K.V ...R..TED. VS    VIE. EANR..QV.D LAN.G.R...
RL3_BACST   MTK.I..R.I ......AE.. DL      HA T.NV.LQK.T IEN.G.E...
RL3_MYCOP    MK.I..R.V E...V.TNS. QL       . L.NT.LQV.T IDS.G.V.V.

RL3_CRYPT   IGYQQVDAKK LTKPQLGHLQ VNNLPPLKHL KEYKVDATHT FTIAQQLDV-
RL3_CYAPA   V..RETKE.N ...A...... KT.NSA.RV. Q.FSIESSDS IEVEKPIT.-
RL3_ECOLI   VTTGAKK.NR V...EA..FA KAGVEAGRG. W.FRLAEGEE ..VG.SIS.-
RL3_BACST   L.FEDISE.R AN...I..AA KA.TA.KRFI R.IRGANINE YEVG.EVK.-
RL3_MYCOP   L.TTDKRVNL VN..E...FK KA.SN.KRFV ..IR--NMQG YE.G.VIN.S

RL3_CRYPT   SIFELGQIVS VSGVSIGKGF AGTVKRHNFT RGPMTHGSK- NHREPGSIGQ
RL3_CYAPA   EL.NDND..N IQ.Y...R.. S.YQ.....A ....S....- ...L.....A
RL3_ECOLI   EL.ADVKK.D .T.T.K.... ......W..R TQDA...NSL S..V......
RL3_BACST   D..SE.D..D .T.I.K.... Q.AI...GQS ....A...RY -..R...M.A
RL3_MYCOP   D..VS.EY.D .T.I.K.... ..GI....YS ....A...GY -..GI..M.A

RL3_CRYPT   GSTPAKVHKG KKMAGRLGGH QVTTKNLTVV HLDKDNNVLV LKGCVPGKRG
RL3_CYAPA   ....GR.YP. TR....K.DS KI.IRG.KI. KV.SERSL.I V..S....P.
RL3_ECOLI   NQ..G..F.. .....QM.NE R..VQS.D.. RV.AER.L.L V..A...AT.
RL3_BACST   -IA.NR.F.T .NLP..M..E R..IQ..KI. KV.PER.L.L I..N...P.K
RL3_MYCOP   -II-NRIF.S ...P.HM.NA KR.IQ..EII AI.QS..IML I..PI..PKN

RL3_CRYPT   NILSIK
RL3_CYAPA   GL.T.TQVKK V
RL3_ECOLI   SD.IV.PAVK A
RL3_BACST   GLVIV.SAVK AKAKAK
RL3_MYCOP   SFVQ..QNVK GMSSKQAVEL LNRNASVQA
```

RPL4

```
RL4_CRYPT   MGSNKIKNLV QYEIQDFVSL NKMDTKSHDS LNLNVSKKSR YLLHRVLTNQ
RL4_BACST        MPKVA L.NQNG-QTV GEIEL-NDAV FGIEPN.H-- -V.FEAVIM.
RL4_ECOLI         ME.. LKDA.S-A-- --LTV-.ETT FGRDFNEA-- -.V.Q.VVAY
RL4_YERSI         ME.. MKDAPG-A-- --LTV-.ETT FGRDFNEA-- -.V.Q.VVAY
RL4_MYCOP         MK.Q VLDTKG-NEI KEIAL-NDYV WGTEPHQQ-- -AIYDTVIS.

RL4_CRYPT   LINNRSGNAC TKTRSEVEGG GKKPWKQKGT GNARAGSSNS PLWKGGGVTF
RL4_BACST   RASM.Q.THK ..N.A..S.. .R...R.... .R..Q..IRA .Q.R...TV.
RL4_ECOLI   AAGA.Q.TRA Q...A..T.S .....R.... .R..S..IK. .I.RS.....
RL4_YERSI   AAGA.Q.TRA Q...A..T.S .....R.... .R.....VK. .I.RS.....
RL4_MYCOP   QAAL.Q.TKK V...A..S.. .R........ .L..Q..IRA .Q....E...

RL4_CRYPT   GPKPR-TFSN KTNKKERLLA LTTALYLKSN NTKVINLDNL DFTNLKTRDL
RL4_BACST   ..V..-SY.Y .LP..V.R.. IKS..SS.VL ENDIVV..Q. SLEAP..KEM
RL4_ECOLI   AAR.Q-DH.Q .V...MYRG. .KSI.SELVR QDRL.VVEKF SVEAP..KL.
RL4_YERSI   AA..Q-DH.Q .V...MYRG. .KSI.SELVR QDRL.IVEKF SVEAP..KL.
RL4_MYCOP   ..T.DINYKK SV...V.A.. FRSV.S..VK ENNLVIV.KF ..AKPS.KEM
```

Figure 3-4 continues

```
RL4_CRYPT     VIKCSNLIEN YKKDQKILFV AEPTASGLWR YVKNISNVDL IYTTGLDLKK
RL4_BACST     .KILN..SVD ----R.A.I. TDELNENVYL SAR..PG.KV VPAN.INVLD
RL4_ECOLI     AQ.LKDMALE ---.--V.II TGELDEN.FL AAR.LHK..V RDA..I.PVS
RL4_YERSI     AQ.LKDMALE ---.--V.I. TGELDEN.FL AAR.LYK..V RDVA.I.PVS
RL4_MYCOP     .VVMK..KID ---...T.I. TKEKEELVVK SSN..TG.KT .SANQ.NVFD

RL4_CRYPT     ILQAHHIIFT CKALNDVKEV FNEQ
RL4_BACST     V.NHDKLVI. KA.VEK.E.. LA
RL4_ECOLI     LIAFDKVVM. AD.VKQ.E.M LA
RL4_YERSI     LIAFDKVVM. AD.VKQ.E.M LA
RL4_MYCOP     L.N.TKLLI. EE.AIA.E.. YA
```

RPL23

```
RL23_CRYPT         MHALI DLVKYPLITD KATRLLELNQ -YTFLTSRVA TKTDIKNAIE
RL23_eugra    MFYFISRKFY .QF..GIL.. .TNK..KN.V -...DVDIQM S.RQF.DL..
RL23_maize             M .GI..AVF.E .SL...GK.. -...NVESGF ...E..HWV.
RL23_ORYSA             M .GI..AVF.E .SL...GK.. -...NVESGF ...E..HWV.
RL23_SINAL             M .GI..AVF.. .SIW..GK.. -...NVESGS .R.E..HWV.
RL23_SPIOL             M .GI..AVF.. .SIQ..GKK. -..SNVESRS .R.E..HWV.
RL23_TOBAC             M .GI..AVF.. .SI...GK.. -..SNVESGS .R.E..HWV.
RL23_ECOLI      MIREER.L KVLRA.HVSE ..STAM.KSN TIVLKVAKD. ..AE..A.VQ
RL23_YERSI      MIREER.L KVLRS.HVSE ..SAAM.K.N TIVLKVAKD. ..AE..A.VQ
RL23_MYCOP          MHIT EVL.K.VL.E .SFAGHKD.V -....VDKK. N.VQ..KTF.

RL23_CRYPT    FLFNVKVISI NTCLLPLKRK RLGKFVGSKP RYKKAVVTLE KNNTINLFSE N
RL23_eugra    TA.S..IT.V .SYVKSS.YY .SNN.E.M.K Y..RMFIK.N DLE..PF..C L
RL23_maize    LF.G...VAV .SHR..G.GR .M.PIL.HTM H.RRMII..Q PGYS.P.LDR ETN
RL23_ORYSA    LF.G...VAV .SHR..G.GR .M.PIL.HTM H.RRMII..Q PGYS.P.LDR EKN
RL23_SINAL    LF.G....AM .SHR..G.V. .M.PIL.HTM H.RRMII..Q PGYS.PPLRK KRT
RL23_SPIOL    LW---NSYEM .SHR..G.GR .M.PIM.HTM H.RRMII..Q SSYS.PPLRK KRT
RL23_TOBAC    LF.G....AM .SHR..G.SR .M.PIM.HTM H.RRMII..Q PGYS.PPLRK KRT
RL23_ECOLI    K..E.E.EVV ..LVVKG.V. .H.QRI.RRS DW...Y...K EGQNLDFVGG AE
RL23_YERSI    K..E.E.EDV ..L.VKG.S. .H.QR..RRS DW...Y...K EGQNLDFIGG AE
RL23_MYCOP    EI.E...E.V R.INYDA.E. ....Y..K.. S....II..K EGQKLDVL.D L
```

RPL2

```
RL2_CRYPT-    MGIRIYKSYT PGTRNRSSSD FVEITKSKPE KSLLRKKLSC A-GRNNRGLI
RL2_CYAPA-    .A..S..A.. ......TI.E .S....E.. ...TFL.HRK K-......I.
RL2_EUGRA-    .I..Y..P.. S...K..V.N .S.....N.. .Y.TFFVHRS K-...S..V.
RL2_EPIFA-    .A.HL..TS. .S...GTV-- -YSQV..N.R .N.IYGQHH. GK...V..I.
RL2_MARPO-    .A..L.RA.. .......VPK .D..V.CQ.Q .K.TYN.HI- KK......I.
RL2_MAIZE-    .AKHL..TPI .S..KGTV-- -DRQV..N.R NK.IHGRHR. GK...A..I.
RL2_SINAL-    .A.HL..TS. .S...GAV-- -DSQV..N.R NN.IYGQHH. GK...A..I.
RL2_ORYSA-    .AKHL..TPI .S..KGTI-- -DRQV..N.R NN.IHGRHR. GK...S..I.
RL2_TOBAC-    .A.HL..TS. .S...GTV-- -DSQV..N.R NN.IYGQHH. GK...A..I.
RL2_ECOLI-    .AVVKC.FTS ..R.HVVKVV NP.LH.G..F AP..E.NSKS G-....N.R.
RL2_BACST-    .A.KK..PTS N.R.GMTVL. .S...TDQ.. ....APLKKR .-....Q.K.
RL2_MYCOL-    .A.KK..PT. N.C..M.V.A .S...TQT.. .R..VSHKDQ .-....Q.K.
RL2_YERSI-    .A.VKC.PTS ..R.HVVKVV NP.LH.G..Y AP..E.LSKS G-....N.R.
```

Figure 3-4 continues

```
RL2_CRYPT-  TVRHKGGGHK QRYRLVDFKR NKLDIPAIVA SVEYDPNRNA RIALLHYQDG
RL2_CYAPA-  .TA.....S. RL..II.... DLKLV..K.. AI........ ........N.
RL2_EUGRA-  .C.TL..... RL..RIE... ...G.LGK.I .I........ ....IY.KN.
RL2_EPIFA-  .T..R..... RL..KIS.IW .EKY.YGRII TI........ Y.C.I..G..
RL2_MARPO-  .SQ.R..... RL..KI..Q. ..KY.TGKIK TI.......T Y.C.IN.E..
RL2_MAIZE-  .A..R..... RL..KI..R. .QK..SGRII TI........ Y.C.I..G..
RL2_SINAL-  ....R..... RL..KI..R. .TK..YGRIV TI........ Y.C.I..G..
RL2_ORYSA-  .A..R..... RL..KI .R. .QK..SGRIV TI........ Y.C.I..G..
RL2_TOBAC-  .A..R..... RL..KI..R. .EK..YGRIV TI........ Y.C.I..G..
RL2_ECOLI-  .T..T..... .A..I..... ..DG...V.E RL......S. N...VL.K..
RL2_BACST-  ....Q..... RQ..II.... D.DG..GR.. TI......S. N...IN.A..
RL2_MYCOL-  ....R...V. RK...I.... ..DN.VGK.. TI......S. N...I..L..
RL2_YERSI-  .T..I..... .H........ ..DG...V.E RL......S. N...VL.K..

RL2_CRYPT-  EKRYILHPKK LAVGDKIYSG INVPIEIGNA MPLYNVPLGT AVHNVELIPG
RL2_CYAPA-  ..G....ARG ....NMV... P.A...V..S L..SEI..A. EI..I..T..
RL2_EUGRA-  D.S..I..FD .C..NN.I.D FFS..K...S L.ISKI.... II....FE..
RL2_EPIFA-  D.......RG AII..TLV.. TE...I.... L..TDM.... .I..I.ITL.
RL2_MARPO-  ......Y.RG IKLD.T.I.S EEA..L...T L..T.M.... .I..I.IT..
RL2_MAIZE-  ........RG AII..T.V.. TK...SM... L..TDM.... .I..I.ITR.
RL2_SINAL-  ........RG AII..T.V.. TE...KM... L..TDM.... .I..I.ITL.
RL2_ORYSA-  ..G.....RG AII..T.V.. TK...SM... L..TDM.... .I..I.ITR.
RL2_TOBAC-  ........RG AII..T.V.. TE...KM... L..TDM.... .I..I.ITL.
RL2_ECOLI-  .R....A..G .KA..Q.Q.. VDAA.KP..T L.MR.I.V.S T.....MK..
RL2_BACST-  .....IA..N .K..ME.M.. PDAD.K.... L..E.I.V.. L...I..K..
RL2_MYCOL-  ......A..G .T..MQ.V.. KEAD.KVA.C LS.M.I.V.. T...I..K..
RL2_YERSI-  .R....A..G .KA..Q.Q.. VDAA.KA..T L.MR.I.V.S T.....MK..

RL2_CRYPT-  RGGQIVRSAG TSAQVVAKDG QVVTIKMPSN EVRMIYKNCY ATIGEVGNAD
RL2_CYAPA-  K...L..... S...LL..E. NY..LRL..G .M.FVR.E.. ....QI...E
RL2_EUGRA-  K....A.A.. .FV.IL.NE. KF...T...G ...LLRRY.W ....Q...L.
RL2_EPIFA-  K...L..A.. AV.KLI..E. KLA.L.L..G ...L.S...S ..V.Q...VG
RL2_MARPO-  K...L..A.. .V.KII..E. .L..LRL..G .I.L.SQK.L ....QI..V.
RL2_MAIZE-  ....LA.A.. AV.KLI..E. KLA.LRL..G ...LVSQ..L ..V.Q...VG
RL2_SINAL-  K...LA.A.. AV.KLI..E. KSA.L.L..G ...L.S...S ..V.Q...VG
RL2_ORYSA-  ....LA.A.. AV.KLI..E. KSA.LRL..G ...LVSQ..L ..V.Q...VG
RL2_TOBAC-  K...LA.A.. AV.KLI..E. KSA.L.L..G ...L.S...S ..V.Q...VG
RL2_ECOLI-  K...LA.... .YV.I..R.. AY. LRLR.G .M.KVEAD.R ..L......E
RL2_BACST-  ....L..A.. .....LG.E. KY.IVRLA.G .....LGK.R ..V.....EQ
RL2_MYCOL-  K....A.... SFC.IISRED KY.LLRLQ.G ..PKVLGT.R .....I..ES
RL2_YERSI-  K...LA.... AYV.I..R.. SY..LRLR.G .M.KVQAD.R ..L......E

RL2_CRYPT-  IKNIRLGKAG RKRWLGIRPS VRGVVMNPCD HPHGGGEGRS PIGRAKPVTP
RL2_CYAPA-  .S..SI.... .N.......T .....K..V. .........A ....ST....
RL2_EUGRA-  HS.VV..... .N....NK.T ....A..... .......... ....P.....
RL2_EPIFA-  VNKKS..R.. S.....K..V ........I. .........A ....K..T..
RL2_MARPO-  VN.L.I.... S.....K..K ........I. .........A ....K..L..
RL2_MAIZE-  VNQKS..R.. S.C...K..V ........V. ........KA ....K..T..
RL2_SINAL-  VNQKS..R.. S.C...K..V ........V. .........A ....K.....
RL2_ORYSA-  VNQKS..R.. S.C...K..V ........V. ........KA ....K..T..
RL2_TOBAC-  VNQKS..R.. S.....K..V ........V. .........A ....K..T..
RL2_ECOLI-  HMLRV..... AA..R.V..T ...TA...V. ..G..H...N -F.-KH....
RL2_BACST-  HELVNI.... .A.......T ...S....V. ........KA ....KS.M..
RL2_MYCOL-  Y.L.NY.... K..F.....T ...SA...N. .........A ....KS.M..
RL2_YERSI-  HMLRV..... AS..R....T ...TA...V. .........N -F.-KH....
```

Figure 3-4 continues

```
RL2_CRYPT-  WGKPALGVKT RRQNKYSDFC IIRSRN
RL2_CYAPA-  .......RR. ..TK....NL ...R.K
RL2_EUGRA-  .......K.. .SPKRF.NKY .....KMV
RL2_EPIFA-  ..Y....RRS .KI.....NF .V.R.SK
RL2_MARPO-  ..H....KRS .KN.....TL .L.R.KNS
RL2_MAIZE-  ..Y....RR. .KRK....SF .L.R.-K
RL2_SINAL-  ..Y....RR. .KRK...ETL .L.R.SK
RL2_ORYSA-  ..Y....RR. .KRK....SF .L.R.-K
RL2_TOBAC-  ..Y....RRS .KR.....NL .L.R.SK
RL2_ECOLI-  ..VQTK.K.. .SNKR-T.KF .V.R.SK
RL2_BACST-  ....T..Y.. .KKKNK..KF ...R.-KK
RL2_MYCOL-  ...K.R.... .DRK.A.NAL ...R.-TK
RL2_YERSI-  ..VQTK.K.. .SNKR-T.KF .V.R.SKK
```

RPS19

```
RS19_CRYPT  MSRSLSKGPY IAAHLLKKLN NVDIQKPDVV IKTWSRSSTI LPNMVGATIA
RS19_CYAPA  .A...K...F ..H.....VE LLNTSGKTE. ......A... ..M...H...
RS19_EUGRA  .....K...F VFYS.I..VD QMNSNRFKS. .L.....C.. I.I.I.N..G
RS19_MARPO  .T..IK...F V.D.....IE .LNLK.EKKI .I....A... V.T.I.H...
RS19_TOBAC  .T...K.N.F V.N.....ID KLNTKAEKEI .V...WA... I.T.I.H...
RS19_SPIOL  .T...K.N.F V.N...R.IE KLNKKAEKEI .V...WA... I.T.I.H...
RS19_ORYSA  .T.K-KTN.F V.H...A.IE K.NMKEEKET .V...WA.S. ..A...H...
RS19_ECOLI  .P...K...F .DL.....VE KAVESGDKKP LR....R... F...I.L...
RS19_MYCOL  .P..VK...I V.S...A.IE KQKNL.NKK. .Q........ T.IF..HK..
RS19_MYCOP  .A...K...F VDEN.F..VT SAKDG---E. ......R... F.EFI.K.FG
RS19_YERSI  .P...K...F .DL.....VE KAVESGDKKP .R....R..V F...I.L...
RS19_BACST   G...K...F CDE..M..IE KLNETGQKQ. ......R... F.QF..H...

RS19_CRYPT  VYNGKQHVPV YISDQMVGHK LGEFSPTRTF RSHIKSDKKA KR
RS19_CYAPA  .H..R..L.. F.T....... ....A..... KG.T...... R.
RS19_EUGRA  .....E.I.. LV....I... ....VQ..NY .G.K.H...T .TKR
RS19_MARPO  .H..QE.L.I ..T.R..... ....A..... .G.--AKNDK .SRR
RS19_TOBAC  IH...E.L.I ..T.S..... ....A..LN. .G.--AKSDN RSRR
RS19_SPIOL  IH..RE.L.I ..T.R..... ....A..LN. WG.--AKNDN .SRR
RS19_ORYSA  IH...E.I.I ..TNP...R. ....V..WH. T.YESAR.DT .SRR
RS19_ECOLI  .H..R..... FVT.E..... ....A....Y .G.AADK.AK .K
RS19_MYCOL  ....RE.I.. ..TEN..... .........Y .G.N.K...I QKK
RS19_MYCOP  .....EFI.. ..TED...N. ....A...K. GG.--G.D.G .KK
RS19_YERSI  .H..R..... FV..E..... ....
RS19_BACST  ..D.RR.... ..TED..... ....A..A.. .G.AGD...T ..
```

Figure 3-4 continues

RPL22

```
RL22_CRYPT                                   MIL SLNSPNVAVP TAKYIRMSPS
RL22_CYAPA                                          MATEVKA I...V.T..Q
RL22_GRATE                                       MT.KTTKIQA .G..V.L.TA
RL22_MARPO                                    MQ TNT.NKKIRA V..H.H...H
RL22_EUGRA                                     M EQKK.LESSA SI..V.I..F
RL22_MAIZE  MTSFKLVKYT PRIKKKK.GL RKLARKVPTD R.LKFERVFK AQ.R.H..VF
RL22_ORYSA  MTSFKLVKYT PRIKKKKSGL RKLARKVPTD R.LKFERVFK AQ.R.H..VF
RL22_tobac                                     M LKKKKTEVYA LGEH.S..AD
RL22_ECOLI                                         METI---A KHRHA.S.AQ
RL22_MYCOL                                         METKKPKA I.RKVSIA.R
RL22_MYCOP                                         MEAK---A KLSM..I..R
RL22_ACHLA                                         MEAK---A IG.T..IA.R
RL22_BACST                                         MQAK---A V.RTV.IA.R

RL22_CRYPT  KIQRVLNQIR GKSYKESLMI LEFMPYAACK PVLQAVQSAG ANAQHNKGIN
RL22_CYAPA  .VR.I.D... ......AV.L .SV...K..S II.KI.D..- ....VT..F.
RL22_GRATE  .TR......K ..K.Q.AIL. ....T.KP.. IIKKILE... N..-L.LKYE
RL22_MARPO  .VR..VS... .R..EQA... ......R..N .I..LLS..A ...N..F.LS
RL22_EUGRA  .VR.I....K .R.A..A... .K....KPST LIFKLLK..V S.S-IK-NYD
RL22_MAIZE  .A....DE.. WRY.E.TV.. .NL...R.SY .I.KL.Y..A ...T.YRDFD
RL22_ORYSA  .V....DE.. WRY.E.TV.. .NL...R.SY .I.KL.Y..A ...T.YRDFD
RL22_tobac  .AR..I.... .R..E.T... ..L...R..Y .I.KLIY..A ...SY.M.SS
RL22_ECOLI  .VRL.ADL.. ..KVSQA.D. .TYTNKK.AV L.KKVLE..I ...E..D.AD
RL22_MYCOL  .ARL.VDL.. ..NIAQAQA. .T.T.KV.AP VI.KLLN..V S..VN.LKL.
RL22_MYCOP  .MRL.ADT.. N.AVSVAVAT .KNLNKD.AE .I.KLLN..V ...VN.N.ME
RL22_ACHLA  .VRL.VDL.. ..NV..AQA. .M.T.RG.SP VIAKVLD..I ..RT..LNL.
RL22_BACST  .ARL.IDL.. ..EVG.RFA. .RHT.K..SP IIEKVLK..V ...E..YDMD

RL22_CRYPT  KNDLVVS--L ASVDNGPVLR RFRPRAQGRG FKIQKFTSHI RIGVQKQVNF
RL22_CYAPA  ..K.II.--K TF..K..T.K .......... Y..L.P.C.. TVQ..D.SL
RL22_GRATE  .QN.IIK--Q .FAND..K.K ..Q......A .R...P.C.. T.NLSIN
RL22_MARPO  .TN.FI.--E IQ.NK.TFFK ..Q....... YP.H.P.C.. T.VLNILPK
RL22_EUGRA  EDAN.LRVLE .RA.A..I.K .LC.H..... .P.K.R.C.. T.---IV
RL22_MAIZE  .TN.FIT--K .E.SRSTIMK K.....R..S YS.K.TMCN. T.VLNIVKKS
RL22_ORYSA  .AN.FIT--K .E.SRSTIMN K.....R..S SP.K.TMC.. T.VLNIVKKS
RL22_tobac  EAN..I.--K .E.NG.TTVK KLK...R..S .P.KRS.C.. T.VMKDISLD
RL22_ECOLI  ID..K.T--K IF..E..SMK .IM...K..A DR.L.R.... TVV.SDR
RL22_MYCOL  REQ.Y.K--E VF.NE.LR.K .MF...K.S. DM.K.R.... TLVITSST.L
RL22_MYCOP  ADK.Y.K--T IF.NE..T.K ......H..A YE.F.R...V V.V.SDEK
RL22_ACHLA  LEN.F.K--E VWANESITMK .ML...K.S. HL.R.R.... TVV.AERE
RL22_BACST  V.N..I.--Q .Y..E..T.K ......M..A SA.N.R.... T.V.SEKKEG

RL22_MAIZE  KWSLDQSKIS KKKRNIIEKY GT
RL22_ORYSA  K
RL22_tobac  DEYVEMYSLK KTRWKKKSTA MPYRDMYNSG GLWDKK
RL22_MYCOL  QTSKEEEQSG SKN
```

Figure 3-4 continues

RPS3

```
RS3_CRYPT   VGQKVNPLGF RLRITSQHRS SWFATKESYP QLLEQDFKIR SYINRELEAA
RS3_CYAPA   ....IH.I.. ..G..QK... C...NPKQ.. T..QE.HL.. Q..EKN.SN.
RS3_EUGRA   M....H.... ..G..KS.S. F.YVERRH.A SFVKE.IV.. NFM.K..LET
RS3_EPIFA   M...I..... ..GT.QS.H. F...QPKN.Y KGIQE.Q... DF.KNYVKNN
RS3_SPIOL   M...I..... ..GT.QS.Y. L..SQPKN.A EG.QE.Q... DC.KNYVQKN
RS3_ORYSA   M...I..... ..G..QN... Y...N-KK.S KVF.E.K... DC.ELYVQKH
RS3_TOBAC   M...I..... ..GT.QG.H. L..SQPKN.S EG.QE.Q... DC.KNYVQKN
RS3_ECOLI   M....H.N.I ..G.VKPWN. T...NTKEFA DN.DS...V. Q.LTK..AK.
RS3_MYCOL   M...T..N.L ..G.IRTWE. Q.CVNDKEI. N.IKE..L.. KL..NFTKKS
RS3_MYCOP   M....S.NVL ..G.VRDWEN R.Y.E.DQ.V KW.D..I... TALFKL.KD.
RS3_ACHLA   M...T..N.L ..G.IRTWE. Q.CVNDKEI. N.IKE..L.. KL..NFTKKS

RS3_CRYPT   --------GI SKIEISRNAN QLEVSVYTSR PGIIVGRSGL GIEKIKTDIL
RS3_CYAPA   --------.. AQ.Y.Q.K.D RI.LELR.A. ..VV...G.R ...VLRKGLK
RS3_EUGRA   --LISLIKIE RIY.F.EQR. NTI.YIHVA. .ERVI..D.Q .LSR.RDILI
RS3_EPIFA   IIISPDTE.. AY...QKRID F.KIMIFIGF KKFLIENRQ. ..--..EALH
RS3_SPIOL   TKTSSGVE.. AR...QKRID LIQ.IIHMGF .KLLIENRPQ .V----E.LK
RS3_ORYSA   IKNSSNYG.. ARV..K.KTD LIQ.EI..GF .ALL.ESR.Q ..----EQLK
RS3_TOBAC   MRTSSGVE.. AR...QKRID LIQ.IIFMGF .KLLIESRPR ..----EELQ
RS3_ECOLI   --SVSRIVIE R------P.K SIR.TIH.A. ...VI.KK.E DV..LRKVVA
RS3_MYCOL   --AISQIDIE RLK.--K.K. RITI..H.AK ..V.I.KD.D TRN.LVAKLK
RS3_MYCOP   --AVSKIDIE R------TTK D.TLFIK.A. .A.VL.QE.K N....VLAVR
RS3_ACHLA   --AISQIDIE RLK.--K.K. RITI..H.AK ..V.I.KD.D TRN.LVAKLK

RS3_CRYPT   R----LLKQD I-SIRINVIE LTNPDADANL IGEFIAQQLE KRVAFRRATR
RS3_CYAPA   D----..GEQ K-Q....... VKQI..E.A. .....T.... R......IV.
RS3_EUGRA   DRMNY..GKT PRI.TCK.VG V.S.NL..R. LADSVRRE.. ..TP.I..MK
RS3_EPIFA   IDLKKNFHYV NRKLI.DI.R I.K.YRNP.I LA....D..K N..S..KTMK
RS3_SPIOL   INVQKE.NCV NRKLN.AITR IAK.YG.P.I LA....G..K S..S..K.MK
RS3_ORYSA   LNVQNI.SSE DRRL.MTL.. IAK.YGEPKI LAKK..LK.. S......TMK
RS3_TOBAC   TTLQKEFHCV NRKLN.A.TR IAK.YGNP.I LA....G..K N..S..K.MK
RS3_ECOLI   D----IAG-- V-PAQ..IA. VRK.EL..K. VADS.TS... R..M....MK
RS3_MYCOL   E----.--TQ K-DVNL..L. VK.S.KI.L. .AQNM.E... N.MF...VQK
RS3_MYCOP   K----TV.NK KLIVNVR... IKS...RST. VARW.GE.IS N.AS..TVQK
RS3_ACHLA   E----.--TQ K-DVNL..L. VK.S.KI.L. .AQNM.E... N.MF...VQK

RS3_CRYPT   QAIQKAQRAN VQGIKVQVSG RLNGAEIARS EWVREGRVPL QTLRANIDYA
RS3_CYAPA   K..TR...RG IE...I.I.. .......... ..S....... .....E...S
RS3_EUGRA   TVMLQ.MK.G AE........ ....I....T ..F....... H....D...F
RS3_EPIFA   K..ELTESED TK..Q..I.. .ID.K....V ..I....... ..IQ.K.N.C
RS3_SPIOL   K..ELTEQ.D TK..QI.IA. .ID.K....I ..I....... ..I..K...C
RS3_ORYSA   K..EL.KKG. IK...I.IA. .........V ..A....... ..I..R.N.C
RS3_TOBAC   K..ELTEQ.D TK..QI.IA. .ID.K....V ..I....... ..I..K...C
RS3_ECOLI   R.V.N.M.LG AK....E... ..G......T ..Y....... H....D...N
RS3_MYCOL   M.....LK.G AK.V.TLI.. ..G....... .GHA...... H....D....
RS3_MYCOP   L..K..LK.G AK...TA... ..G.V.M..T .GYL.CS... S...N.....
RS3_ACHLA   M.....LK.G AK.V.TLI.. ..G....... .GHA...... H....D....
```

Figure 3-4 continues

```
RS3_CRYPT    TKEAHTTYGI GIKVWVFNG EQTPTYAVI
RS3_CYAPA    Y.R.Q.I.   ..V...I.K. .VI.GNPTEI SE
RS3_EUGRA    NDI...I..V .......YKV
RS3_EPIFA    SYMVR..H.V ....I.I.IE KE
RS3_SPIOL    AYTVR.I..V ....I.I.M. .E
RS3_ORYSA    YYA.Q.I..V ......I.QD .E
RS3_TOBAC    SYTVR.I..V .. .I.I.LD .E
RS3_ECOLI    .S.......V I.V...I.K. .ILGGM.AVE Q....PEKPA
RS3_MYCOL    AV.......V ......I.H. .VL.GQTILD TRKPFASQSS NTPNRRPRNF
RS3_MYCOP    LY..P....Q I.V...INH. .VF------. .KK..ERMNN
RS3_ACHLA    AV.......V ......I.H. .VL.GQTILD TRKPFASQSS NTPNRRPRNF

RS3_MYCOL    KGGNNNHVNA KKN
RS3_ACHLA    KGGNNNHVNA KKN
```

RPL16

```
RL16_CRYPT   MLSPKRTKFR KPHRGRLRGI ATRGNTLIFG DYGLQALEPI WLSXRQIEAT
RL16_CYAPA   ....R..... .QQ...MK.. S....N.V.. .F.......A .ITS.....S
RL16_GRATE   .......... .Q..N.MN.K .SK...IA.. E.A..T...V ..TA......
RL16_EPIFA   ....QK.R.. .Q....MK.. SY...NIC.. K...K....A .ITP.....G
RL16_MARPO   .......... .Q.C.N.K.. S....VIC.. KFP......S .ITS.....G
RL16_CHLSP   .......... .....H...K ......IV.. .FA...Q..C .ITS.....G
RL16_CHLRE   .......... .....H...K .....KIV.. .FA...Q..C .ITS.....G
RL16_EUGRA   .......... .Y.....T.K IY--DKVV.. N.A..S...G .ITS.....A
RL16_MAIZE      ....R.. .Q....MK.K SC...HIC.. R.A..V...A .ITA.....G
RL16_ORYSA   .......R.. .Q....MK.K SY...CIC.. R.A......T .ITA.....G
RL16_TOBAC   .......R.. .Q....MK.. SH...HIS.. K.A......A .ITS.....G
RL16_SPIOL   .......R.. .Q....MK.. SY...RIC.. R.A......A .ITS.....G
RL16_ECOLI   ..Q....... .M.K..N..L .-Q.TDVS.. SF..K.VGRG R.TA.....A
RL16_MYCOP   ..Q.....Y. ....VSYE.K .KGAKEIN.. EF..M..DGA .IDNH....A

RL16_CRYPT   RRTITRQVKR VGRLWIRVFP DKSISAKPPE TRMGAGKGAP EYWVAVIKPG
RL16_CYAPA   ..A.N.Y.R. G.KI...I.. ..PVTMR.A. ....S..... .....IV...
RL16_GRATE   ......Y... G.KI...... ..P.T.R.A. ....S....T ..........
RL16_EPIFA   ..A...KFR. G.KI.V.... ..PVTVRSS. ....S...SH K.F...V...
RL16_MARPO   ..A...YAR. G.K....I.. ..P.TIR.A. ....S...S. ......V...
RL16_CHLSP   ..VL..Y.R. G.K....I.. ..AVTMR.AG ....S..... D.....VH..
RL16_CHLRE   ..VL..Y.R. G.K....I.. ..AVTMR.AG ....S..... D.....VH..
RL16_EUGRA   ..V...YA.. G.K....I.. ..PVTFRAA. ....S...NV .....IV...
RL16_MAIZE   ..AM..YAR. G.KI.V.I.. ..PVTIR.T. ....S...S. ......V...
RL16_ORYSA   ..AM..YAR. G.KI.V.I.. ..PVTIR.T. ....S...S. ......V...
RL16_TOBAC   ..AM..NAR. G.KI.V.I.. ..PVTLR.A. ....S...S. ......V...
RL16_SPIOL   W.AM..NGR. G.KI.V.I.. ..PVTVR.A. ....S...S. ......V...
RL16_ECOLI   ..AM..A... Q.KI...... ..P.TE..LA V...K...NV .....L.Q..
RL16_MYCOP   .IAM..YM.. D.KI.M.I.. HMAMTK..A. V...S...N. .K....V.K.
```

Figure 3-4 continues

```
RL16_CRYPT  HILFEINGVS QDLRYLAFKN ASYKLPIKTK FISR
RL16_CYAPA  RVI....... .EMAKA..RL .TF....... ...SRV
RL16_GRATE  ......A... KQTAQE.M.L .......... ..TK
RL16_EPIFA  L..Y..G..T ENIAKW.ILI .AS.M.MQ.Q ..ISG
RL16_MARPO  K..Y..S... ENIARA.M.I .A..M..R.Q ..TTSSLNKK QEI
RL16_CHLSP  K..Y.MQ... ETIARQ.MRI .A..M.V... .LTKQA
RL16_CHLRE  K..Y.MQ... ETIARQ.MRI .A..M.V... .LTKTV
RL16_EUGRA  K..Y.VL.I. ESIAKYSL.I .G..M....R V.VKI
RL16_MAIZE  R..Y.MS... ETVARA.ISI .AS.M..RSQ .LRLEI
RL16_ORYSA  R..Y.MG... ETVARR.ISI .AS.M..RSQ .LRLEI
RL16_TOBAC  R. Y.MG..T ENIARR.ISL .AS.M..R.Q ..IS
RL16_SPIOL  R..Y..S..A ENIARRDVAI .AS.M..R.Q ..ISG
RL16_ECOLI  KV.Y.MD..P EE.ARE...L .AA......T .VTKTVM
RL16_MYCOP  T.M..VAQ.N EQVARE.LRL .MH....RC. .VK.GEN
```

RPL29

```
RL29_CRYPT  MTTNLDST-Q LEKLTDTDIN DTVLKLKKEL FELR-LQKAT RQEIKPHLFK
RL29_BACSU  .KA----N-E IRD..TAE.E QK.KS..E.. .N..-F.L.. G.LENTARIR
RL29_ECOLI  .KAK-----E .REKSVEEL. TEL.N.LR.Q .N..-M.A.S G.LQQS..L.
RL29_CHLTR  .GAKKNLLAE .REKSSEELD EFIRDN..A. .A..AEAALQ NKVV.T.Q.S
RL29_HALMA  ..V-.HVQ-E IRDM.PAERE AELDD..T.. LNA.AV.A.G GAPEN.GRI.

RL29_CRYPT  QKKKLIAKLL TIKSKKS
RL29_BACSU  EVR.A..RMK .VIREREIAA NK
RL29_ECOLI  .VRRDV.RVK .LLNE.AGA
RL29_CHLTR  LY..S..RA. I..QE.KGRV HG
RL29_HALMA  ELR.A..RIK ..QGEEGDLQ ENE
```

RPS17

```
RS17_CRYPT  MSIKER---- --LGLVISDK MDKTVVVSIA NRVTHKRYGK IVTKTKKYKV
RS17_CYAPA  .AS...---- --V.V..RNP QE...I.AVN ...R.NK.S. .IIR....Q.
RS17_BACSU  ..E--.NQRK VYQ.R.V... ....IT.VVE TYKK.TL... R.KYS..F.A
RS17_ECOLI  .TD--KI--R TLQ.R.V... .E.SI..A.E RF.K.PI... FIKR.T.LH.
RS17_CHLTR  .ASDV.GRRK TKI.V.V.S. .E.....RVE RVYS.PQ.A. V.RDSS..YA
RS17_MYCOP  .-Q--.NSRR VLI.K.V... ....IT.LVE TYKN.PI.K. R.KYS....A
RS17_THERM  .------PKK VLT.V.V... .Q...T.LVE RQFP.PL... VIKRS...LA

RS17_CRYPT  HDPNNNCQVG DLILINETRP LSKTKRWMFK EIKQ-KSLKL DKDTIGE
RS17_CYAPA  ..HSHI.KL. .EVK.S.VK. I......IIS .VLS-STVNP E.FGD
RS17_BACSU  ..E..QAKI. .IVK.M.... ..A...FRLV .VVE-EAVII
RS17_ECOLI  ..E..E.GI. .VVE.R.C.. .....S.TLV RVVE-.AVL
RS17_CHLTR  .NELDV-KE. .TVR.Q.... .......RVV GRVN
RS17_MYCOP  ..E.QVA.M. .KVE.M.... .....NFRLV RVIEKAT.
RS17_THERM  ...EERYK.. .VVE.I.A.. I..R..FRVL RLVE-EGRLD LVEKYLVRRQ

RS17_THERM  NYASLSKRGG KA
```

Figure 3-4 continues

RPL14

```
RL14_CRYPT  MIQTQTYLTV ADNSGAKKIM CIRILGG-N- RKYASIGDVI IGVVKDATPN
RL14_CYAPA  ...P.S...A ......R.L. ...V...-GN .R..R..... VA....GI..
RL14_MARPO  ...P....N. ......R.L. ...VI.T-SN ....N...I. .A...E.V..
RL14_eugra  ..KP....KI ...T..Q... ......P-.C -Q..N...I. .A...E.I..
RL14_Chlre  ..KPLS..N. ......REL. ...A...-SY .ES.N..... .A.....L..
RL14_maize  ...P..L.N. ......R.L. ...VI.AAGN QR..R..... .A.I...V.Q
RL14_ORYSA  ...P..L.N. ......R.L. ...VI.AASN QR..R..... VA.I...V.Q
RL14_SPIOL  ...P..H.N. ......REL. ....I.A-SN .R..R..... VA.I.E.I..
RL14_TOBAC  ...P..H.N. ......REL. ....I.A-SN .R..H..... VA.I.E.V..
RL14_ECOLI  ...E..M.N. ......RRV. ..KV...SH- .R..GV..I. KITI.E.I.R
RL14_BACSU  ...QE.R.K. . ....REVL T.KV...SG- ..T.N..... VCT..Q...G
RL14_BACST  ...QESR.K. ......REVL V.KV...SG- .R..N....V VAT......G
RL14_CHLTR  ...QESQ.K. ...T....VK .FKV...SR- .R..TV.... VCS.R.IE.D
RL14_THERM  ...P....E. ...T..R... ...V.K.S.- A...TV.... VAS..E.I.R

RL14_CRYPT  MPVKRSDVVR AVIMRTKNTI RRKDGMSIRF DDNAAVII-N KENNPRGTRV
RL14_CYAPA  I.I.K..T.K ...V..RKEL K.DN..N.C. ........-. ADG.......
RL14_MARPO  ..I.K.EI.. ...V..CKEF K.NN.SI.K. ......V.-. Q.G..K....
RL14_eugra  .V..K..I.K ...V..VKGV ..ES..A... .E......-. NDRS.K...I
RL14_Chlre  .......I.. ...V..RKG. ..EN..A... ........-. ..G.......
RL14_maize  ..LE..E.I. ...V..RKEF KGD..II..Y ........DQ .G-..K....
PL14_ORYSA  ..LE..E.I. ...V..CKEF KCE..II..Y ........DQ .G-..K....
RL14_SPIOL  T.LE..E.I. ..V...CKEL K.DN..I..Y ......V.ID Q.G..K...I
RL14_TOBAC  ..LE..E... ...V..CKEL K.DN..I..Y ......V.DQ EGRKSK...I
RL14_ECOLI  GK..KG..LK ..VV...KGV ..P..SV... .G..C.LLN. NSEQ.I...I
RL14_BACSU  GV..KGE..K ...V...SGA ..S..SY.S. .E..C...RD DKS-.....I
RL14_BACST  GV..KGQ..K ..VV...RGV ..P..SY... .E..C...RD DKS-.....I
RL14_CHLTR  SS..KG...K ...V..R.D. H....STL.. .T.SC...DD .G.-.K...I
RL14_THERM  GA..EG...K ..VV...KEV K.P..SA... ........N. QLE-......

RL14_CRYPT  FGPIARELRD KDFTKIVSLA PEVL
RL14_CYAPA  ...V...... .N....I... ....
RL14_MARPO  .........E SN........ ....
RL14_eugra  .........E .E.V..M... ...V
RL14_Chlre  .......... .N........ ....
RL14_maize  ..AV.E...E LNL....... ....
RL14_ORYSA  ..A..E...E LN........ ....
RL14_SPIOL  ..A......Q -K.A...... ....
RL14_TOBAC  ..A......E LN........ ....
RL14_ECOLI  ...VT....S EK.M..I... ....
RL14_BACSU  ...V.....E NN.M...... ...I
RL14_BACST  ...V...... ...M..I... ...I
RL14_CHLTR  ...V...I.. RG.V..S... ...I
RL14_THERM  ...V.....E .G.M...... ....
```

RPL24

```
RL24_Chlre  MAAMV...LQ SSFTSLSLSS NSFLGQR.LF PSPTTLQVKT EGHS...PCL
RL24_SPIOL  MAAMV...LQ SSFTSLSLSS NSFLGQR.LF PSPTTLQVKT EGHS...PCL
RL24_TOBAC  M.....AALQ SSFAGLS... TSFFGQR..F SPPLSLPPLV ..KSTEGPCL
RL24_PISUM  MVAMAMASLQ SSMSSLSLSS NSFLGQP..L SP.ITLSPFL QGKPTEKKCL
```

Figure 3-4 continues

```
RL24_CRYPT                             MTIKQGD KVQVIAGSYK GEITEVLKVI
RL24_BACST                             .HV.K.. .....S.KD. .KQGVI.AAF
RL24_BACSU                             .HV.K.. ..M..S.KD. .KQGTI.AAF
RL24_ECOLI                      M AAK--.RRD. E.I.LT.KD. .KRGK.KN.L
RL24_CHLTR                        MKRRSVCV.. T.Y.L..ND. .KQGK..RCL
RL24_THERM                      M QAKVHV.K.. T.L.AS.K.. .RVGK.KA L
RL24_Chlre    IVMRIKRWER KDCKPNSLPK LHKRHV.V.. T.K..S.GE. .K.G.IS.IH
RL24_SPIOL    IVMRIKRWER KDCKPNSLPK LHKRHV.V.. T.K..S.GE. .K.G.IS.IH
RL24_TOBAC    IQAKLKRWER KECKPNSLPV LHK.HV.L.. T.KI.S.HD. .KVG.ITEI.

RL24_CRYPT    RKSNSLILKN INIKNKH-VK PKKEGEVGQI KQFEAPIHRS NVMLYDEESQ
RL24_BACST    P.K.RV.VEG V..VK..A-. .SQANPQ.G. IEK.....V. K..PL.PKTG
RL24_BACSU    P.KDRVLVEG V.MVK..S-. .TQANPQ.G. SNQ.....V. ...PL.PKTG
RL24_ECOLI    SS-GKV.VEG ..LVK..QKP VPALNQP.G. VEK..A.QV. ..AIFNAATG
RL24_CHLTR    --KDKVVVEG ..VRV.N.I. RSQ.NPK.KR INI...L.I. ..R.--SIDN
RL24_THERM    PRKMAV.VEG V.LVK.-A.R .SPKHPQ.GF VEQ...L.A. K.RPICPACG
RL24_Chlre    KHNSTV.I.D L.F.T..-.. S.E...Q... IKI..A..S. ....ILK.QE
RL24_SPIOL    KHNSTV.I.D L.F.T..-.. S.E...Q... IKI..A..S. ....ILK.QE
RL24_TOBAC    KHNSKVVV.D V.L.T..-.. SRS.D.P... VKI.....S. .....SK.QK

RL24_CRYPT    IRSRSKFIIS QDGKKVRVLK KLVKN
RL24_BACST    EPT.IGYK-I V......YA. .SGEILDK
RL24_BACSU    EVT.VGYK-V E.......A. .SGQVLDK
RL24_ECOLI    KAD.VG.RF- E......FF. SNSETIK
RL24_CHLTR    QPA.LFVKVT EK.RELWNKH SDGSSSLYRL VRERKG
RL24_THERM    KPT.VRKKFL E..R.I.ACA .CGGSLDVEE
RL24_Chlre    VAD.VGHK.L E.VR...Y.I .TGEIVDTPD RWKE.IQNKK ESETAVAVAA
RL24_SPIOL    VAD.VGHK.L E.VR...Y.I .TGEIVDTPD RWKE.IQNKK ESETAVAVAA
RL24_TOBAC    VA..VGHKTL DN..R..Y.I .TGEIIDSAE NWKKAVKEKE KTTEAVAAAS
```

RPL5

```
RL5_CRYPT       MNKSLKEI YYQDVIPGLI EQFNYTNIHQ VLKITKITLN RGLGEASKNN
RL5_CYAPA        VQR..:' .E.E..KQ.M TR.Q.K...E .P.LK...V. ......AQ.A
RL5_eugra        MQR..SF .LETI..K.K .E.G.V.SYR .P.LK..VI. ..FD.SCQ.S
RL5_astlo        MQK..S. .ITK.C.I.V NE.L...FFE IP..N.VVIS ..F..SCNSS
RL5_BACSU        MNR...K .NKEIA.A.M TK...DSVM. .P..E..VI. M.V.D.VQ.A
RL5_ECOLI        MAK.HDY .KDE.VKK.M TE...NSVM. .PRVE..... M.V...IADK
RL5_CHLTR                                  M. IPVLK..VIS M..A..A.DK
RL5_THERM     MPLDVA..KK ..DE.R.E.. RR.G.Q..WE .PRLE.VV1. Q.....KEDA

RL5_CRYPT     KILEASIKEF ELISGQHPLI NKARKSVAGF KIREGMPVGI SVTLRKKLMY
RL5_CYAPA     ......V..I TE.T..KAIV TR.K.AI... .L.QD..I.V M....GDY..
RL5_eugra     ....VLLN.L .I....K.I. S..K.AI.N. .LK.K....M FL...SEK..
RL5_astlo     ....SLLV.L KN....K.IL C.SKN.ISN. .VKK...I.M F...HGDK..
RL5_BACSU     .AIDSAVE.L TF.A..K.VV TR.K..I... RL.....I.A K....GER..
RL5_ECOLI     .L.DNAAADL AA....K... T......... ...Q.Y.I.C K....GER.W
RL5_CHLTR     NLFQ.HLE.L AV....K..V TR.KN.I... .L...QGI.A K....GIR..
RL5_THERM     R...KAS..L A..A..K.A. TR.K..ISN. .L.K...I.L R....GDR.W
```

Figure 3-4 continues

```
RL5_CRYPT   TFLEKLIHLS LPRIRDFRGV SVKSFDGRGN YNLGIKEQLI FPEIEYDQVD
RL5_CYAPA   A..DR..N.. .........I TA........ ....L..... ...VD..SIE
RL5_eugra   S..DR..N.. .......Q.I NKNC...S.. FSF.LS..SM ....NF.KMI
RL5_astlo   S..DR..N.. F..M...N.L NI.G...F.. ..V.LS..S. ......SSIL
RL5_BACSU   D..D...SV. ...V...... .K........ .T........ ....D..K.T
RL5_ECOLI   E.F.R..TIA V........L .A........ .SM.VR..I. ....D..K..
RL5_CHLTR   D.MDRFCNIV S........F .C.G-....C .S..LDD.Q. ...VDL.A.K
RL5_THERM   I.....LSVA .........L NPN....... ....L..... ....T..M..

RL5_CRYPT   QVRGLDISIT TTAKTQQEGI ALLRALGMPF NDN
RL5_CYAPA   .I..M....V .....N...L ...KS..... AES
RL5_eugra   K.Q..N.T.V ...E.N..AF F..KE..I.. R.
RL5_astlo   KNK.MN.T.V .....DL.SF S..KG..F.. CV
RL5_BACSU   K...M..V.V ...N.DE.AR E..TQV.... QK
RL5_ECOLI   R..... T.. ....SDE..R ...A.FDF.. RK
RL5_CHLTR   RSQ.MN.TWV ...Q.DA.CL T..ECM.LR. KKAQ
RL5_THERM   VP..M..AVV .....DE.AK ...EL..F.. RK
```

RPS8

```
RS8_CRYPT   VT-NDTVSDM LTRVKNANLA KHQVVQVPAT KMTKSIAHVL LEEGFIESIE
RS8_CYAPA   .V-...IA.. ..GI...... ..K.AR.K.. .I.RCL.N.. K...L.QNF.
RS8_ASTLO   M.TI..I..V I..I....IL .LDR.ELIN. .VAIG.C.I. KDR...N.FG
RS8_EUGRA   M.NI.VI... ...I..SL.I .ARK.N.IN. .L.VN..EI. KK....D.F.
RS8_EPIFA   MG-R..ILEI INSI...DRG RKR..RITS. NI.ENFVKI. FI.....NAR
RS8_MARPO   MG-...IAN. I.SI.....G .IKT...... NI.RN..KI. FQ....DNFI
RS8_ORYSA   MG-K..IA.L ..SI...DMN .KGT.W.VS. NI.EN.VKI. .R......VW
RS8_MAIZE   MG-K..IA.L ..SI...DMN .KGT.R.VS. NI.EN.VKI. .R......VW
RS8_TOBAC   MG-R..IAEI I.SI...DMD RKR..RIAS. NI.EN.VQI. .R.....NVR
RS8_ECOLI   MSMQ.PIA.. ...I..GQA. NKAA.TM.SS .LKVA..N.. K......DF-
RS8_MYCOL   M.-T.VIA.. ...I....QR YLKT.S..SS .VKLE..RI. K.....SDFT
RS8_THERM   M-LT.PIA.. ...I...TRV YKESTE...S RFKEE.LKI. AR....KGY.

RS8_CRYPT   EVGLDI---- -----NRQLL LSLKYKGRE- -REPVINA-- --LKRISRPG
RS8_CYAPA   .IENNL---- -----QNE.. I......KK- -.Q.I.T.-- --.....K..
RS8_ASTLO   .FLNSSDRMS NRRFIQKYII VN.....ERR ------SPCI KE.R...K..
RS8_EUGRA   --LA.ATCLT ENGVIKKYIT IF.....PKQ ------VSYI TKI..V.K..
RS8_EPIFA   KHREKN---- -----KYYFT .T.RH--.RN SKR.Y..IL- -N.....W..
RS8_MARPO   DNKQNT---- -----KDI.I .N...Q.KK- -KKSY.TT-- --.R...K..
RS8_ORYSA   KHQESN---- -----RYF.V ST.RHQK.KT RKGIYRTRT- -F........
RS8_MAIZE   KHQESN---- -----RYF.V ST.RHQR.KT RKGIYRTRT- -F........
RS8_TOBAC   KHREKN---- -----KYF.V .T.RH--.RN RKR.YR.IL- -N........
RS8_ECOLI   K.EG.T---- -----KPE.E .T...FQGKA --------VV ESIQ.V....
RS8_MYCOL   -.EG.V---- -----KKTIN IE...Q.KT- -.--..QG-- --..K..K..
RS8_THERM   R.EV.G---- -----KPY.R IH...GP.RQ GPD.RPEQVI KHIR......
```

Figure 3-4 continues

```
RS8_CRYPT   LRVYANRKFL PRVLGGLGIA VISTSKGVLT DTKARTQGLG GEVLCYIW
RS8_CYAPA   ..G...H..  .......... IL...S.IM. .QT..HK.C. ........
RS8_ASTLO   R...VGY.N. HKTK..IELF .L.....LI. .YT..EK.I. ..L.FS.C
RS8_EUGRA   ..T.SSY.R. QS.A..V.LT .L.....LM. .RL..SNKI. ..I.F...
RS8_EPIFA   ..I.S.SQQI .LI...I..V ILY..R.IM. .RE..LK.I. ..L.....
RS8_MARPO   ..I.S.H..I .K....M..V IL...R.IM. .RE..QKKI. ..L...V.
RS8_ORYSA   ..I...YQGI .K....M... IL...R.IM. .RE..LNRI. ........
RS8_MAIZE   ..I...YQGI .K....M... IL...R.IM. .RE..LNRI. ........
RS8_TOBAC   ..I.S.YQRI ..I...M..V IL...R.IM. .RE..LE.I. ..I.....
RS8_ECOLI   ..I.KRKDQ. .K.MA..... .V.. ...M. .RA..QA... ..II..VA
RS8_MYCOL   .....QAN.I .Q..N....S IV...Q.IM. GK...LANA. ....AF..
RS8_THERM   R...VGV..I ...RR..... IL..P..... .RE..KL.V. ..LI.EV.
```

RPL6

```
RL6_CRYPT   MSRIGKLPVK FSEKVTMKI- DQDNIIVKGP KGELALGLSK NINITIENNT
RL6_CYAPA   ......RLIN IPSQ..VS.- KDQVFS.... ....SKQIPY G.QVVQQ.DH
RL6_ECOLI   ...VA.A..V VPAG.DV..- NGQV.TI..K N...TRT.ND AVEVKHAD..
RL6_MYCOL   .....NRLLQ IPNG.EV..A ENNLVTIT.S ..T.SKQF.P L.K.EV.E.K
RL6_MYCOP   .....NRLLQ IPNG.EV..A ENNLVTIT.S ..T.SKQF.P L.K.EV.E.K
RL6_CHLTR   ...KARD.IV LPQG.EVS.- QN.E.S.... ..S.TQV.A. EVE.AVKG.E
RL6_BACST    ..V..K.IE IPAG..VTV- NGNTVT.... ....TRTFHP DMT..V.G.V
RL6_THERM   .....R..IP VPKG.QVQV- SPGLVK.... ....SVPV.P ELKVVV.E.V

RL6_CRYPT   LFVKPVTKEP QVL-KLFGTY RAIINNMVVG VTKGFEKRLE LQGVGYRAQL
RL6_CYAPA   .V.ER.AESL LAR-..H.LC .TLVS.L.Q. .FQ...R... I.......M
RL6_ECOLI   .TFG.RDGYA DGWAQA-..A ..LL.S..I. ..ED.T.K.Q .V......AV
RL6_MYCOL   .IT.RLNEQK HTK-Q.H..T NSLLQG.LT. .SE..K.E.Q IT....K.AV
RL6_MYCOP   .IT.RLNEQK HTK-Q.H..T NSLLQG.LT. .SE..K.E.Q IT....K.AV
RL6_CHLTR   V..A.AAHVV DRPGRMQ.L. W.L.A...K. .HT....... MI...F..AV
RL6_BACST   IT.TRPSD.K HHRA-.H..T .SLLA...E. .S..Y..A.. .V......SK
RL6_THERM   VR.ERPSD.R RHRS-.H.LT .TL.A.A.K. .SE.YV.E.L IK.I....R.

RL6_CRYPT   QGKDLSLSVG YSHPVVIKAP TGINIAVENN TIVIISGISK ELVGQIASNI
RL6_CYAPA   D..K.V.NI. F......EP. .E.QLQ.... .NI..K..D. ....KL.AE.
RL6_ECOLI   K.NVIN..L. F....DHQL. A..TAECPTQ .EIVLK.AD. QVI..V.ADL
RL6_MYCOL   N.SK.N..L. .....EFEI. D.VV.QAVKP .ELA.T..D. Q....V.A..
RL6_MYCOP   N.SK.N..L. .....EFEI. D.VV.QAVKP .ELA.T..D. Q....V.A..
RL6_CHLTR   ..SL.D..I. V...TKMPI. ..LEVS..K. .LIS.K..N. Q...EF.ACV
RL6_BACST   ...K.V.... .....E.EPE E.LE.E.PSQ .KI.VK.AD. QR..EL.A..
RL6_THERM   V.RSIE.T.. F.....VEP. E..TFE.PEP .RI

RL6_CRYPT   RSIKPPEPYK GKGIRYVGEF VRKKAGKAGK K
RL6_CYAPA   .AVR...... ......L..N .KR.V..... .
RL6_ECOLI   .AYRR..... ...V..AD.V ..T.EA.KK
RL6_MYCOL   .AYRK..... ....K.KN.T IIR.E...AG .
RL6_MYCOP   .AYRK..... ....K.KN.T IIR.E...AG .
RL6_CHLTR   .AKR...... ......EN.Y ..R.....A. TGKK
RL6_BACST   .AVR...... ......E..L ..L.E..T..
```

Figure 3-4 continues

RPL18

```
RL18_CRYPT                        MKRT NKIKGTLERP RLSVFRSNCH IYAQVIDDSS
RL18_CYAPA      MKLTRKQ ATQRRHRRVR R.VF..S... ..A....HQ. ....I...TQ
RL18_ECOLI        MDKKS ARIRRATRAR R.LQE-.GAT ..V.H.TPR.  .....APNG
RL18_MYCOL      MKFTKTE ARKRRHFRVR H.VV..A... ..N..K..TN F...I...TK
RL18_CHLTR   MESSLYKKTS GKARRALRVR KAL..CSLK. ....VKT.K. V.V.L...VE

RL18_CRYPT   GMTIVSTSTL DKDVK-SLLN NTSTCEASKI VGQVIAKKTL ARNIKQVIFD
RL18_CYAPA   HR.LAAS... EPS..N.E.S S....A..A. ...L....A. EKG.T..V..
RL18_ECOLI   SEVL.AA..V E.AI-AEQ.K Y.GNKD.AAA ..KAV.ERA. EKG..D.S..
RL18_MYCOL   .V.L..A... KM.L.----- SK.NIQ.AEK .AEELT..A
RL18_CHLTR   .K.LAFI... A.VA.T.G.T RKNQ-DNA.A L.IK..ELGK GLQVDR.V..

RL18_CRYPT   RGKRVYHGRI SALAEAARES GLEF
RL18_CYAPA   ..GKI....V RT.......A ..Q.
RL18_ECOLI   .SGFQ....V Q...D....A ..Q.
RL18_CHLTR   ..AHK...VV AMV.DG...G ..Q.
```

RPS5

```
RS5_CRYPT    MLNAKKSNKT KEKETDWQER VIQVRRVTKV VKGGKKLSFR AIIILGNERG
RS5_CYAPA    .A.RQ.MS.. RD.KP..... .V.I...S.. .......... ..VVI.....
RS5_ECOLI    .AHIE.---- --QAGEL..K L.A.N..S.T ....RIF..T .LTVV.DGN.
RS5_CHLTR        TMLSRN SH..DQLE.K .LV.N.CC.. ....R.F..S .L.LV.DRK.

RS5_CRYPT    QVGVGVGKAS DVIGAVKKAV TDGRKNLINI PLTNQNSIPH IVQGYSGAAK
RS5_CYAPA    .....I.... ...N.....A A..K.HVVEV ...RS..... PID.IG...R
RS5_ECOLI    R..F.Y...R E.PA.IQ..M EKA.R.M..V A.-.NGTLQ. P.K.VHTGSR
RS5_CHLTR    RL.F.FA..N ELTD.IR.GG DAA....VS. NSLEGG.... E.LVNHDG.E

RS5_CRYPT    VIIKPSAPGS GVIAGGSVRT ILELAGIKNI LAKQLGSSNP LNNARAAANA
RS5_CYAPA    ..MR...E.T ......A... V.....VR.. . .....N.. .......M..
RS5_ECOLI    .FMQ.ASE.T .I....AM.A V..V..VH.V ...AY..T.. I.VV..TIDG
RS5_CHLTR    LLL..AK..T .IV..SRI.L ...M..V.D. V..S...N.. M.QVK..FK.

RS5_CRYPT    LINLRTYTSV LNDRNLDLH
RS5_CYAPA    .SR.K.FSQF AK..GVIAE
RS5_ECOLI    .E.MNSPEM. AAK.GKSSVE EILGK
RS5_CHLTR    .LT.SCKDDI MKR.AVIND
```

Figure 3-4

Table 3-1. Intergenic regions in the *Guillardia theta S10/spc* operon-like gene cluster and the location of putative ribosome binding sites

Intergenic region		Ribosomal binding site	Distance from start codon (bp)
Genes	Size (bp)		
rpl3	\	AGAA	9 (ATG)
rpl3-rpl4	23	AGGAA	43 (ATG)
rpl4-rpl23	10	AGGAG	28 (ATG)
rpl23-rpl2	15	N/D	(ATG)
rpl2-rps19	32	N/D	(ATG)
rps19-rpl22	26	N/D	(ATG)
rpl22-rps3	29	GGAG	5 (GTG)
rps3-rpl16	52	N/D	(ATG)
rpl16-rpl29	10	N/D	(ATG)
rpl29-rps17	23	AGGAG	7 (ATG)
rps17-rpl14	5	GGAG	10 (ATG)
rpl14-rpl24	2	AGAGG	10 (ATG)
rpl24-rpl5	9	AAGAA	25 (ATG)
rpl5-rps8	27	AAGAG	6 (GTG)
rps8-rpl6	37	AAGAG	8 (ATG)
rpl6-rpl18	10	AGG	20 (ATG)
rpl18-rps5	27	AGAA	33 (ATG)
rps5-secY	72	AGAA	31 (ATG)

Table 3-1

Table 3-2. The percentage of identity of ribosomal proteins from the chloroplast genome of *Guillardia theta* compared to that in other organims. The percentage identity of ribosomal protein sequences between *Guillardia theta* (Crypt) and *Cyanophora paradoxa* (Cyapa), *Marchantia polymorpha* (Marpo), *Euglena gracilis* (Eugra), *Oryza sativa* (Orysa), *Nicotiana tabacum* (Tobac), *Escherichia coli* (Ecoli), *Bacillus stearothermophilus* (Bacst) are listed. Numbers represent the percentage of identity between *Guillardia theta* and that species; '\' indicates that the data are not available.

Crypt_RP	Cyapa	Marpo	Eugra	Orysa	Tobac	Ecoli	Bacst	Therm	Chltr
RPL3	52	\	\	\	\	45	43	\	\
RPL4	\	\	\	\	\	30	31	\	\
RPL23	\	\	37	39	38	31	\	\	\
RPL2	67	60	63	55	58	53	61	\	\
RPS19	70	58	58	48	51	57	60	\	\
RPL22	54	49	38	39	42	37	41	\	\
RPS3	59	\	44	46	40	47	\	\	\
RPL16	67	64	57	56	58	57	\	\	63
RPL29	\	\	\	\	\	32	34	\	29
RPS17	44	\	\	\	\	42	43	37	38
RPL14	71	70	68	66	65	57	61	69	59
RPL24	43	\	\	\	45	32	\	30	28
RPL5	62	\	51	\	\	56	56	56	44
RPS8	64	50	48	48	45	48	\	47	35
RPL6	54	\	\	\	\	41	51	40	49
RPL18	64	\	\	\	\	41	\	\	39
RPS5	66	\	\	\	\	38	\	\	38

Table 3-2

and only 9 are present in chloroplasts of tobacco and liverwort. Ribosomal protein genes in the plastid genome of *Guillardia theta* are arranged in the same order as those in *E. coli* and other organisms, indicating that these genes and their organization are evolutionarily related. Genes for ribosomal proteins Rps10, Rps14, Rpl30 and Rpl15 are absent in the plastid genome of *Guillardia theta* as well as in other plastid genomes, suggesting that they may have been the first ones transferred to the nucleus (and/or nucleomorph) following the endosymbiotic event. Interestingly, the *S10* ribosomal protein gene is missing from the plastid ribosomal protein gene clusters, which may indicate that the regulation of transcription and translation of the plastid gene cluster is different from that of *E. coli*. In *Guillardia theta*, the *rps10* gene has been found elsewhere on the plastid genome (Douglas, 1991).

Although the *Guillardia theta* plastid ribosomal protein genes and their organization resemble the *E. coli S10* and *spc* ribosomal protein operons, there are several differences. First, the *rpS10* gene is missing. In *E. coli*, the RNA sequence around the start codon of this gene bears important signals for the feedback regulation of the operon at the translational level; thus absence of this gene from *Guillardia theta* plastid genome indicates that the regulation of gene expression in this region may be very different from that in *E. coli*. Second, the intergenic spacer between *rps17* and *rpl14* is only five bp, unlikely long enough to accommodate a promoter to allow transcription initiation to take place. In *E. coli*, the intergenic spacer between *rps17* and *rpl14* is 163 bp, which separates these two genes into two operons and contains regulatory signals for the transcription as well as translation of *spc* operon.

To further investigate the expression of ribosomal protein genes encoded in this region in the *Guillardia theta* plastid genome, Northern blot analysis was performed. In Northern blot analysis, five probes were used. As shown in Fig. 3-6, these probes cover

Figure 3-5. Comparison of ribosomal protein gene organizations. Ribosomal protein gene content and organization are compared between the *S10-spc* operons of *Escherichia coli* and corresponding plastid gene clusters of *Guillardia theta, Cyanophora paradoxa, Euglena gracilis, Nicotiana tabacum,* and *Marchantia polymorpha.* Each box represents a ribosomal protein gene. The length of each box does not reflect the size of the protein it encodes. Solid lines between two boxes represent intergenic spacers, but their lengths do not reflect the lengths of the intergenic sequences. In the gene names, S indicates that the gene product is assembled into the small ribosomal subunit, L indicates that the gene product is assembled into the large ribosomal subunit.

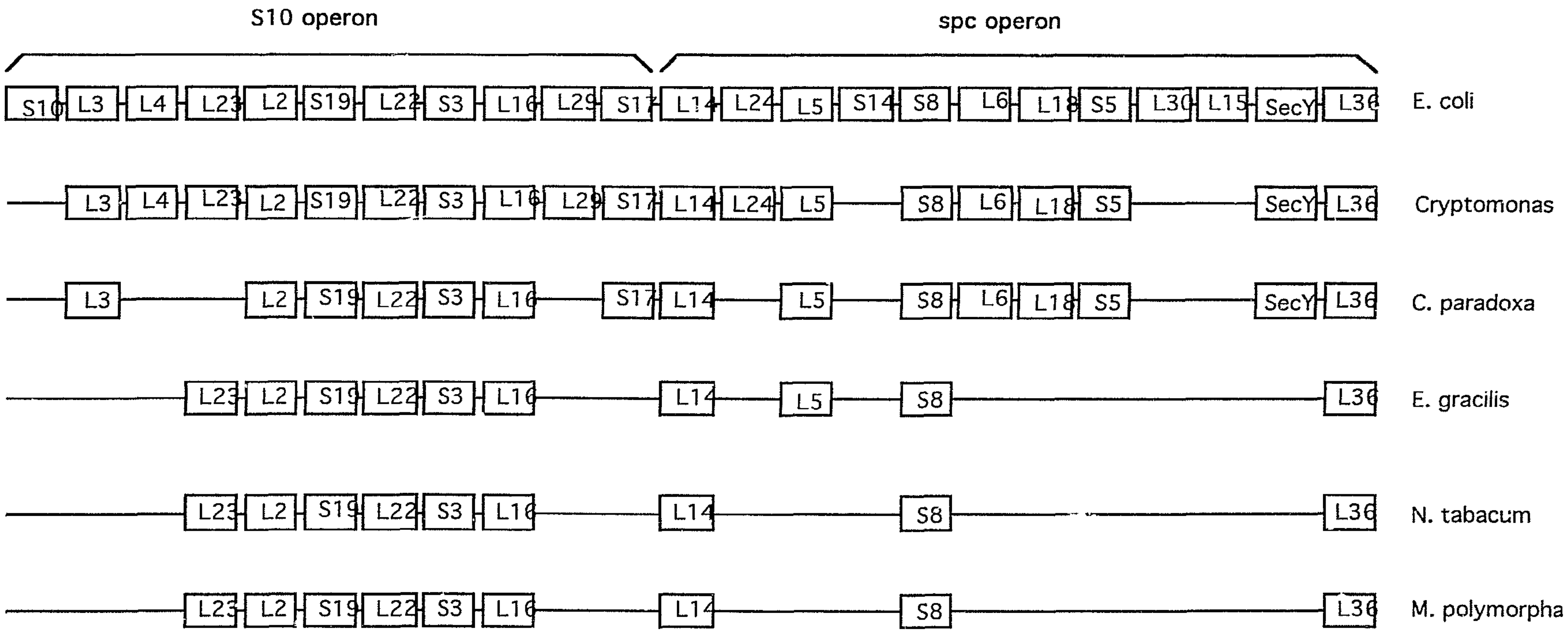

S10 operon
spc operon
S10 L3 L4 L23 L2 S19 L22 S3 L16 L29 S17 L14 L24 L5 S14 S8 L6 L18 S5 L30 L15 SecY L36
E. coli
L3 L4 L23 L2 S19 L22 S3 L16 L29 S17 L14 L24 L5 S8 L6 L18 S5 SecY L36
Cryptomonas
L3 L2 S19 L22 S3 L16 S17 L14 L5 S8 L6 L18 S5 SecY L36
C. paradoxa
L23 L2 S19 L22 S3 L16 L14 L5 S8 L36
E. gracilis
L23 L2 S19 L22 S3 L16 L14 S8 L36
N. tabacum
L23 L2 S19 L22 S3 L16 L14 S8 L36
M. polymorpha

Figure 3-5

almost the entire seauenced region. Northern blot analyses show that all the five probes hybridize to a 10 kb mRNA as shown in Fig. 3-7, suggesting that genes found in the sequenced region are co-transcribed as a single transcript. Probes A, B, C and E also hybridized to our other bands besides the 10 kb band. These four bands comigrate with the most predominant and the only visible bands on an EtBr-stained gel; therefore they are likely non-specific signals. The 10 kb band is invisible on EtBr-stained gel, but this signal in the Northern blot is much stronger than the other four bands, suggesting that it is the authentic mRNA signal.

The DNA sequence downstream of the *secY* and *rpl36* genes (Douglas, 1992) to the genes of *str* operon (Douglas, 1991) has also been determined (Wang, Liu and Douglas, unpublished data). Between the *rpl36* gene and *rps12* are *rps13, rps11, rpoA, rpl13, rps9*, and an *orf72*, in the order of *S13-S11-rpoA-L13-S9-orf72. S13, S11* and *rpoA*, which encodes for the alpha subunit of RNA polymerase, are the three genes of alpha operon in *E. coli*. In *E. coli*, the alpha operon is adjacent to the *spc* operon and contains 5 genes (*S13-S11-S4-rpoA-L17*; Nomura, 1984)). However, the *rpl13* and *rps9* genes, which form the *L13* operon, are about 260 kb away from the *S10, spc* and *alpha* operons (Fig. 3-8).

Figure 3-6. Preparation of DNA probes used in Northern blot analysis. Each open box represents a gene identified by DNA sequencing. The length of the open box reflects the relative size of the protein it encodes. Restriction sites used in the preparation of DNA probes are shown in the middle. A short solid line above the probe's name reflects the size of the probe and the genes it encodes. Probe A is a 1.1 kbp Xba I-Bgl II fragment, corresponding to a stretch encompassing amino acid residue 1 of Rpl3 to amino acid residue 148 of Rpl4; probe B is a 1.4 kbp Bgl II-Bgl II fragment, corresponding to amino acid residue 148 of Rpl4 to amino acid residue 3 of Rps19; probe C is a 1.5 kbp Bgl II-Hind III fragment, corresponding to amino acid residue 3 of Rps19 to amino acid residue 36 of Rpl16; probe D is a 0.6 kbp Pst I-Hind III fragment, corresponding to amino acid residue 83 of Rpl14 to amino acid residue 44 of Rpl5; probe E is a 2.0 kbp Sal I-Sal I fragment (the insert of clone S-7), corresponding to amino acid residue 82 of Rps8 to amino acid residue 106 of SecY.

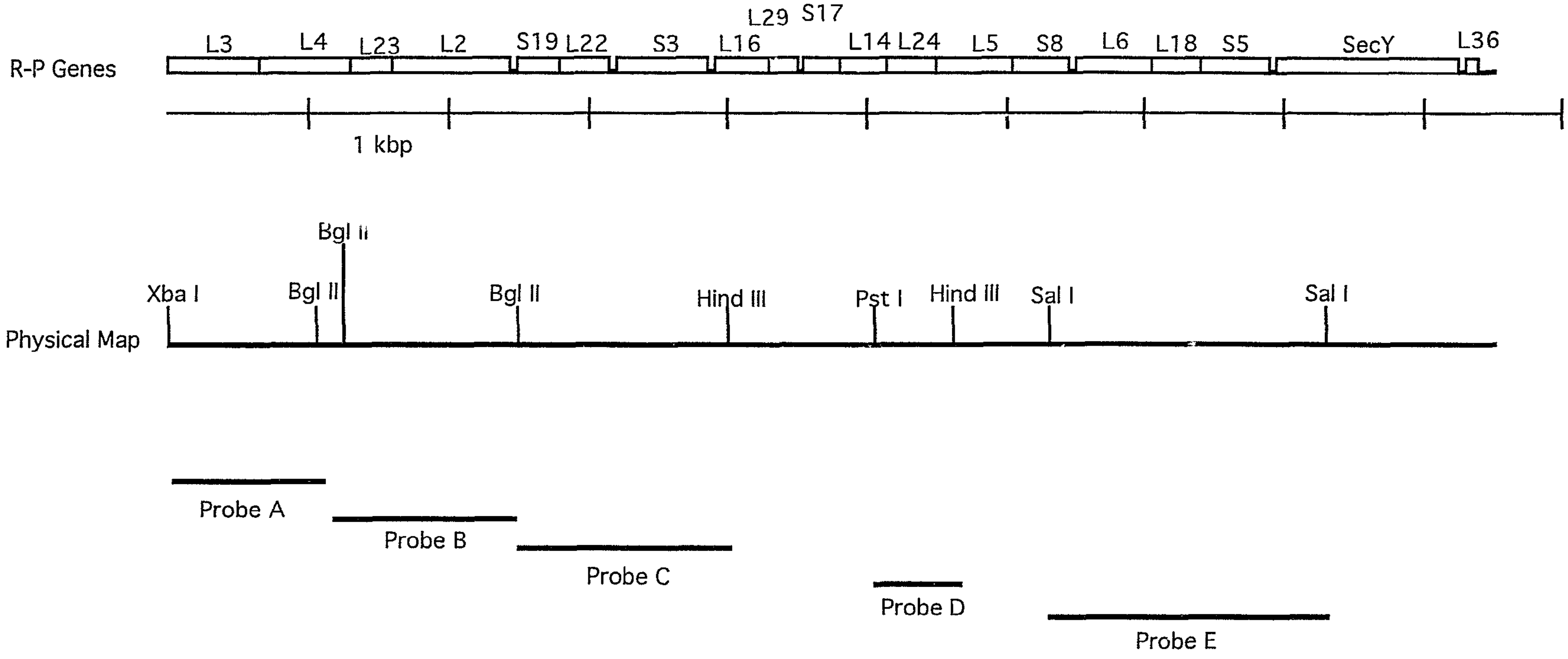

L29 S17
L3
L4
L23
L2
S19
L22
S3
L16
L14
L24
L5
S8
L6
L18
S5
SecY
L36
R-P Genes
1 kbp
Bgl II
Xba I
Bgl II
Bgl II
Hind III
Pst I
Hind III
Sal I
Sal I
Physical Map
Probe A
Probe B
Probe C
Probe D
Probe E

Figure 3-6

Figure 3-7. Northern hybridization with gene-specific probes. Lanes A, B, C, D and E are the Northern blot hybridized with probes A, B, C, D and E, respectively. The *dnaK* lane was hybridized with *dnaK* gene-specific probe, which is located immediately upstream of ribosomal protein gene cluster. For each lane, 5 μg of *Guillardia theta* total RNA was resolved in an agarose gel and blotted to a nylon membrane as described in the Materials and Methods section. One lane was stained with EtBr along with the RNA markers (BRL). RNA size markers are 9.49, 7.46, 4.4, 2.37, 1.35 and 0.33, and shown aside.

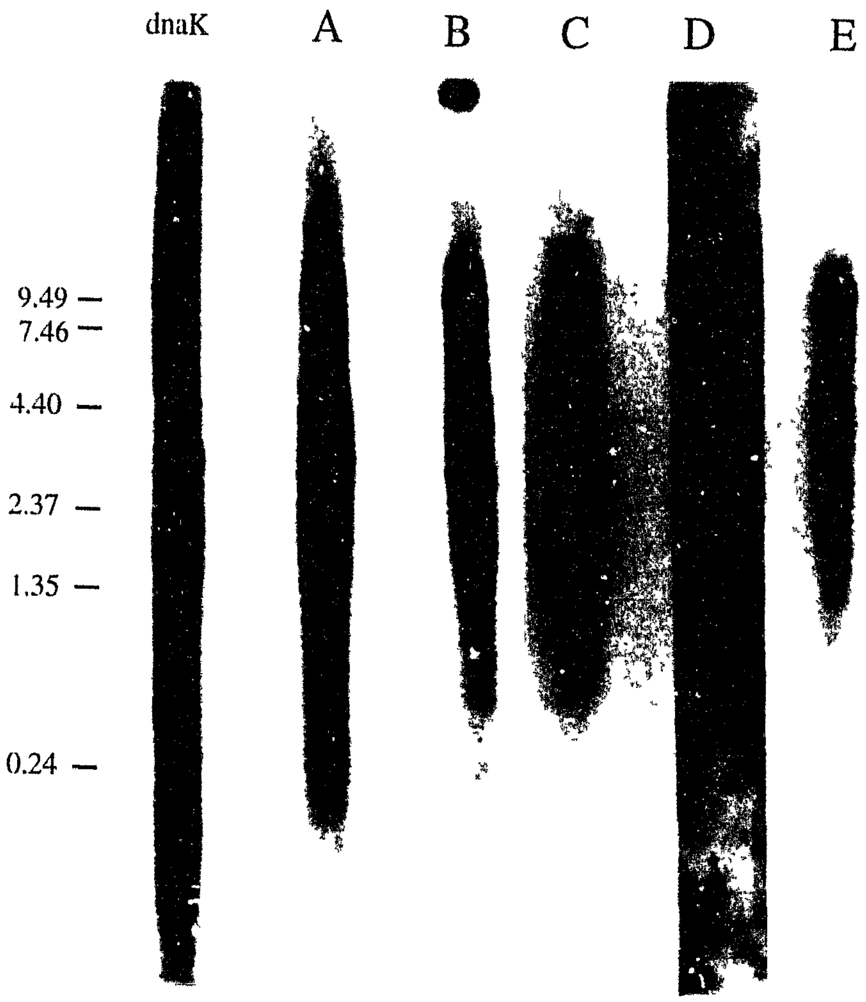


Figure 3-7

Figure 3-8. Rearrangement of ribosomal protein gene clusters in the plastid genome of *Guillardia theta phi* relative to the corresponding operons of *E. coli*. Each open box represents a ribosomal protein gene operon, with the name of the operon in the box. The length of the box does not reflect the length of the operon. Solid thick arrow indicates the direction of transcription and translation of that operon. The Arabic number beneath each box is the number of genes encoded in that operon.

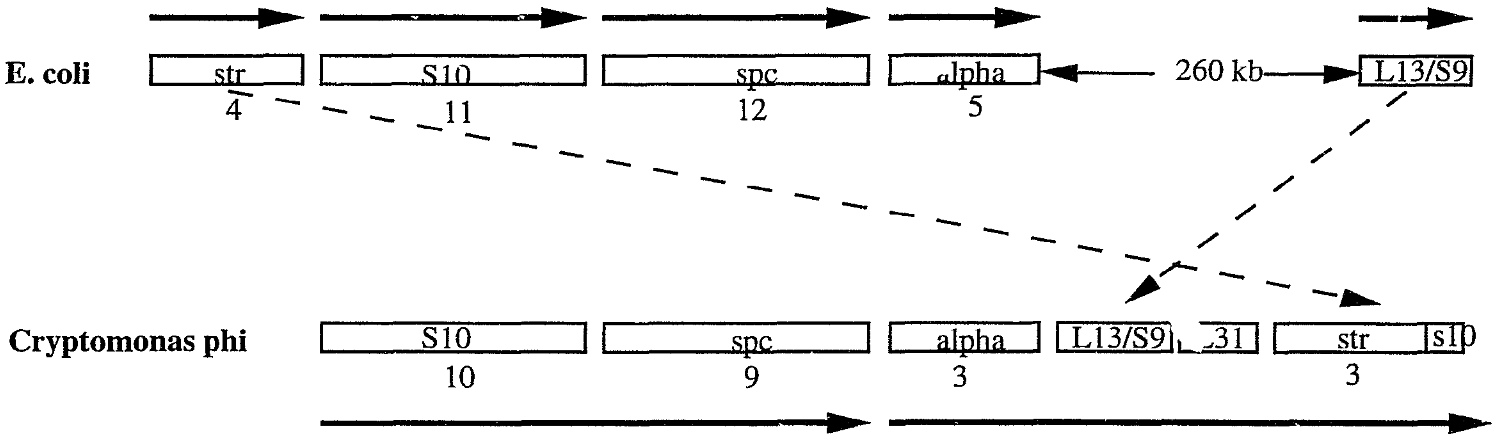

E. coli
str
S10
spc
alpha
260 kb
L13/S9
4
11
12
5
Cryptomonas phi
S10
spc
alpha
L13/S9
L31
str
s10
10
9
3
3

Figure 3-8

DISCUSSION

Guillardia theta is unique among photosynthetic organisms. Unlike other photosynthetic organisms such as plants, *Guillardia theta* obtained its plastid through secondary endosymbiosis. Evidence supporting the secondary endosymbiotic origin of the plastid has come from the observation that the plastid is surrounded by four layers of membranes, that it contains a nucleomorph in the periplastidal space between the two pairs of membranes (Greenwood, 1974), and that it has two phylogenetically distinct 18S rRNAs (Douglas et al., 1991). Because of these unique features, study of gene content and gene organization in the plastid genome of this organism would be interesting and informative in understanding the evolution of this algae and possible impact of secondary endosymbiosis.

Physical mapping of the plastid genome of *Guillardia theta* revealed that it is a circular double-stranded DNA genome of 118 kb in size (Douglas, 1991). The complete sequence of 8.3 kbp DNA fragment from the plastid genome of *Guillardia theta* described here revealed that the plastid genome is highly AT-rich, with 67% A+T versus 33% G+C. The high AT content in this genome is mainly due to the strong codon usage bias towards A and T at the third position. Besides these common features of a typical plastid genome (size, GC content, and codon usage bias), the plastid genome of *Guillardia theta* uses GTG (valine) as the start codon for a few genes. In the *S10/spc* operon-like gene cluster described here, GTG is used as start codon for *rps3* and *rps8* genes. The assignment of GTG as start codon for these genes is based the evidence that there is no in-frame ATG codon upstream of these genes, that protein sequence alignment indicates protein starts at this position, and that a putative ribosome binding sequence is identified upstream of that position. Whether there is a tRNA recognizing GTG as a start codon in this genetic system, or the GTG is changed into ATG through RNA editing, remains unknown.

The complete DNA sequencing identified 18 genes encoded in the 8.3 kbp region of the *Guillardia theta* plastid genome. Comparison of protein sequences deduced from these genes with homologous protein sequences from other organisms revealed that they are ribosomal protein genes whose counterparts in *E. coli* are encoded in the *S10* and *spc* operons. Overall, the *Guillardia theta* protein sequences are more similar to the corresponding sequences of *Cyanophora paradoxa* (43-71%), less similar to the corresponding sequences of *Marchantia polymorpha* (49-70%), and least similar to the corresponding sequences of the other organisms (28-68%). Among the different ribosomal proteins, Rpl14 is the most highly conserved (57-71%), while Rpl24 is the least well conserved (28-45%). The alignment of Rpl24 protein sequences from different groups also shows that *Guillardia theta* Rpl24 is similar in size to bacterial Rpl24, whereas the Rpl24 of plant chloroplast contains up to 23 extra amino acid residues at the N-terminus and 25 amino acid residues at the C-terminus. Although the sequence conservation between Rpl4 of *Guillardia theta* and other organisms is higher than that of Rpl24, small gaps were found in the N- and C-terminal regions of Rpl4.

The arrangement of ribosomal protein genes in the *Guillardia theta* plastid genome is identical to that in other organisms and in the *E. coli S10* and *spc* operons, indicating that plastids of different groups are derived from a common ancestor. Significantly, *Guillardia theta* contains most of the genes found in the eubacterial *S10* and *spc* operons, significantly more than what is found in other plastids. Furthermore, the *S10* and *spc* operon-like gene cluster in the plastid genome of *Guillardia theta* is most similar to the corresponding gene cluster in the cyanelle of *Cyanophora paradoxa*. Among the 23 genes encoded in the *S10* and *spc* operons of *E. coli*, 19 are encoded in the gene cluster of *Guillardia theta* plastid, compared to 15 in the gene cluster of cyanelle of *Cyanophora paradoxa,* 10 in *Euglena gracilis*, and only 9 in plant and liverwort chloroplasts. More than that, protein sequence similarity is higher between *Guillardia theta* and *Cyanophora paradoxa* than between *Guillardia theta* and other organisms. GTG is the start codon of *rps3* and *rps8* genes in both *Guillardia theta* and *Cyanophora paradoxa,* whereas ATG is the start codon of these genes in all the other organisms. While it has been suggested that cyanelle of *Cyanophora paradoxa* is the intermediary between cyanobacteria and chloroplast (Christine et al., 1990), the plastid of *Guillardia theta* may be the intermediary between cyanobacteria and cyanelle. It was noticed that in *Porphyra purpurea* (Reith and Munholland, 1993) and *Odontella sinensis* (Kowallik et al. 1995) plastid genomes, the genes and their organization in the equivalent gene cluster is almost identical to that in *Guillardia theta*, upstream of that gene cluster is also the *dnaK* gene encoded on the opposite strand, suggesting that these organisms are derived from a more recent common ancestor.

The large *S10/spc* ribosomal protein operon is highly conserved among the plastid genomes of *Guillardia theta,* cyanelle, *Euglena* and higher plants, with *Guillardia theta* having retained most genes, cyanelle having retained fewer genes, *Euglena* having retained fewer genes still, and plants having retained the least number of the genes. The arrangement of genes is identical in *Guillardia theta* cyanelle, *Euglena* and higher plants. In comparison to *E. coli*, *rps10*, *rps14, rpl30* and *rpl15* are absent in *Guillardia theta,* cyanelle, *Euglena* and higher plants, indicating that these genes must have been among the first ones to have been lost following the endosymbiotic event. The genes *rpl4, rpl29* and *rpl24*, which are absent in cyanelle, *Euglena* and plants, are present in *Guillardia theta* whereas *rpl3, rps17, rpl6, rpl18, rps5* and *secY* are present in both *Guillardia theta* and cyanelle, but are absent in *Euglena* and plants. Interestingly, *rpl23* is present in *Guillardia theta*, *Euglena* and higher plants, but is absent in cyanelle. The *Guillardia* *theta rps10* has been found at the 3'-end of *str* operon (Douglas, 1991), which has been located about 3 kbp downstream of the *S10/spc* operon. In the cyanelle of *Cyanophora paradoxa*, *rps10* has also been found at the 3'-end of *str* operon, but the *str* operon is located 2.5 kbp upstream of the *S10/spc* operon (Neumann-Spallert et al., 1990).

A 10 kb transcript was detected by Northern hybridization with probes covering all the sequenced region, suggesting that all the genes encoded in the sequenced region are co-transcribed. Although little is known about promoter sequences determining transcription in the plastid genome of *Guillardia theta*, the 10 kb transcript very likely starts a short distance upstream of *rpl3*, since another gene is located just 208-nucleotides upstream of *rpl3* and encoded on the opposite strand. In addition to the 10 kb transcript, probes A, B, C, and E all hybridized to four other bands ranging between 4.4 and 1.35 kb. These four bands most likely represent non-specific artifacts, although the possibility that they may represent processed mRNAs has not been ruled out. First, the four bands co-migrated with four predominant bands that are visible on EtBr-stained gels and believed to be rRNAs. Second, the amount of RNA in these four bands is much larger than that of the 10 kb transcript, but the 10 kb transcript produced a Northern hybridization signal that was same or higher than that produced by the four bands. Probe D, which is relatively short in size, hybridized only to the 10 kb transcript. Third, if these four bands were the processed mRNA, they would have been hybridized by one probe or the other but not by all the four probes spanning the whole region. Notably, the equivalent gene cluster in cyanelle of *Cyanophora paradoxa* is also co-transcribed as a single mRNA. Besides the single mRNA, RNA species with smaller sizes were observed, those smaller RNA species form different hybridization patterns with different probes, indicating a limited processing of the full length mRNA (Christine et al., 1990). In *E. coli*, the intergenic space between *rps17* and *rpl14* is 163 bp, which separates these two genes into two operons (*S10* and *spc*), whereas in *Guillardia theta*, the intergenic spacer between these two genes is only 5 bp, unlikely big enough to accommodate a separate promoter.

The arrangement of ribosomal protein genes in operon-like gene clusters in the plastid genome of *Guillardia theta* is identical to that found in *E. coli*; however, a major genomic rearrangement has occurred in the plastid genome. The *str* operon, which is upstream of *S10* and *spc* operons in *E. coli*, has been relocated to a position downstream of the alpha operon in the *Guillardia theta* plastid genome. The *rps10* gene, which is the first gene in the *S10* operon, has been relocated along with the *str* operon. The *L13/S9* operon, which is found 260 kbp away from the other ribosomal protein operons in *E. coli*, has been inserted between the alpha operon and the relocated *str* operon. The *rps4* gene, which is the third gene of the *alpha* operon in *E. coli*, has been relocated to a position adjacent to the *rbcL* gene (Douglas and Durnford, 1990).

Northern hybridization analysis described here detected a 10 kb mRNA that appears to cover the *S10/spc* operon, while Douglas (1991) detected a 7 kb mRNA that appears to cover the *alpha, L13/S9* and *str* operons. Thus the 5 operons which are separately transcribed in *E. coli*, may have been fused into two operons in *Guillardia theta* plastid. The unusually large ribosomal protein operons of *Guillardia theta* represent a major departure from the popular view that organellar operons have become much reduced relics of their prokaryotic form and are evolving toward extinction. Clearly, these operons appear to be evolving in the opposite direction , as judged by the observation that they contain most of the genes found in *E. coli* (19 out of 23 in the case of *S10* and *spc* operons), additional genes (*rpl13* and *rps9*) have become incorporated into the gene cluster, and individual operons are fused into large hybrid operons (*S10/spc*, *alpha/L13/S9/str*).

Chapter IV The discovery of three novel genes, the *acpA, hlpA* and *dnaK* genes in the plastid genome of *Guillardia theta*

INTRODUCTION

The endosymbiosis hypothesis suggests that contemporary plastids and mitochondria evolved from free-living eubacteria that entered early eukaryote cells as endosymbionts (Gray and Doolittle, 1982). This hypothesis predicts that subsequent to endosymbiosis, a massive loss of the endosymbiont's genetic materials and gene transfer to the nucleus occurred. The retention of a few genes in the endosymbionts' genomes resulted in the organellar genomes we see today, which accounts for the much smaller size of plastids and mitochondria genomes relative to free-living eubacteria. So far, a few plastid and mitochondrial genomes have been completely sequenced, and their gene contents determined. In the cases of plastid genomes, plant chloroplast genomes contain about 60 identified protein-encoding genes and about 30 unidentified open reading frames, in addition to tRNA and rRNA genes. The identified genes encode proteins involved in transcription (RNA polymerase subunits), translation (ribosomal proteins, translation factors), photosynthesis, and probably chlororespiration (*ndh* genes). Various chloroplast genomes studied so far show similar gene contents, while several exceptions have been found in which a gene is present in the chloroplast genome of one organism but absent from that of another. Such exceptions have proven useful in studying the process of gene transfer from chloroplast to nucleus (Baldauf et al., 1990).

Most of the studied plastid genomes are circular double-stranded DNA molecules of 100 to 200 kbp in size. Among them, rhodophyte plastid genomes tend to be larger in size, ranging from 175 to 195 kbp. Chromophyte (including cryptomonad) plastid genomes tend to be smaller, ranging from 115 to 135 kbp. The sizes of chlorophyte and metaphyte plastid genomes are somewhere in between, except that *Chlamydomonas* chloroplast genomes are between 200 and 300 kbp in size. Although the sizes of plastid genomes are within a relatively narrow range (compared to the size range of mitochondrial genomes), rhodophyte and chromophyte plastid genomes contain a substantially larger number of genes than chlorophyte and metaphyte plastid genomes (Reith, 1995). The larger coding capacity of rhodophyte and chromophyte plastid genomes is accounted for the absence or near absence of introns and the much smaller intergenic spacers in their genomes. Among the genes found only in the plastid genomes of rhodophytes and chromophytes, a majority encode proteins involved in photosynthesis and gene expression, but a few belong to other categories. The size of the plastid genome of *Guillardia theta* is 118 kbp, which is close to the low end of the plastid genome size range. However, previous studies of ribosomal protein operons in the plastid genome of *Guillardia theta* revealed that it contains more genes in the *S10/spc* operon than any other plastid genome characterized to date.

In this Chapter, I describe the discovery in the *Guillardia theta* plastid genome of three genes that had not been found in any other plastid genomes before. Complete DNA sequence determination of a 2.8-kbp DNA fragment of this genome revealed three open reading frames (*orf81*, *orf93*, and *orf627*). Analyses of the protein sequences deduced from these open reading frames revealed that *orf81* encodes an acyl carrier protein, *orf93* encodes a histone-like protein, and *orf627* encodes an Hsp70-like heat shock protein. The *orf81*, *orf93*, and *orf627* were subsequently named as *acpA*, *hlpA*, and *dnaK* genes, respectively. All three genes are located immediately upstream of the *S10-spc* ribosomal protein operons, with the *dnaK* gene encoded on the opposite DNA stand relative to all the other genes. These three genes are described for the first time in an organellar genome, and each of them represents a new functional class of organellar genes.

RESULTS

A. Cloning and DNA sequence determination of a 2.8-kbp DNA fragment isolated from the *Guillardia theta* plastid genome

The BS7 plasmid clone described previously in Chapter III contains a 12-kbp BamH I-Sal I DNA fragment from the plastid genome of *Guillardia theta*. This plasmid DNA was cut with EcoR I and Xba I, and a 2.8 kbp-EcoR I-Xba I DNA fragment was isolated and cloned into plasmid vector pUC118 (Fig. 4-1). The complete sequence of this DNA fragment was determined on both strands by the dideoxy termination method.

Analysis of the DNA sequence revealed that it has three open reading frames that were subsequently named *acpA*, *hlpA*, and *dnaK* genes, respectively (Fig. 4-1). All three genes are located upstream of the ribosomal protein genes described in Chapter III, and the *dnaK* gene is encoded on the opposite DNA strand relative to the other genes (see Fig. 4-1). The complete DNA sequence and the translation of the three genes are shown in Fig. 4-2. The intergenic spacer between *acpA* and *hlpA* is 105 bp long, and the intergenic spacer between *hlpA* and *dnaK* is 13 bp long. Neither intron nor potential hairpin-forming sequences were found. All three genes use ATG as the initiation codon. The *acpA* and *hlpA* genes use TAA as the termination codon, while *dnaK* uses TGA as its termination codon. The GC content and codon usage bias in this sequence are similar to that of the downstream ribosomal protein genes.

B. Identification of the acyl carrier protein gene (*acpA*)

The *acpA* gene predicts a polypeptide (AcpA) of 81 amino acid residues long, which contains 17 acidic amino acid residues, 4 basic amino acid residues, and 26 hydrophobic amino acid residues. AcpA has a calculated molecular mass of 8.9-kDa.

Figure 4-1. Gene map of the 2.8-kbp EcoR I-Xba I DNA fragment of *Guillardia theta* plastid genome. The top line represents a portion of the BS7 DNA fragment described in Chapter III. The *S10-spc* ribosomal protein operons are located downstream of the Xba I site, as the arrow indicates. The EcoR I and Xba I sites used in isolating the 2.8 kbp DNA fragment are shown. Number in parenthesis indicates the number of nucleotides between that site and the BamH I site. Three genes, *acpA*, *hlpA* and *dnaK*, were identified within this region. The *acpA* and *hlpA* genes (gray boxes) are transcribed left to right, same as are the downstream ribosomal protein genes. The *dnaK* gene (black box) is encoded on the opposite DNA strand and transcribed right to left. Length of each box reflects the relative size of the gene.

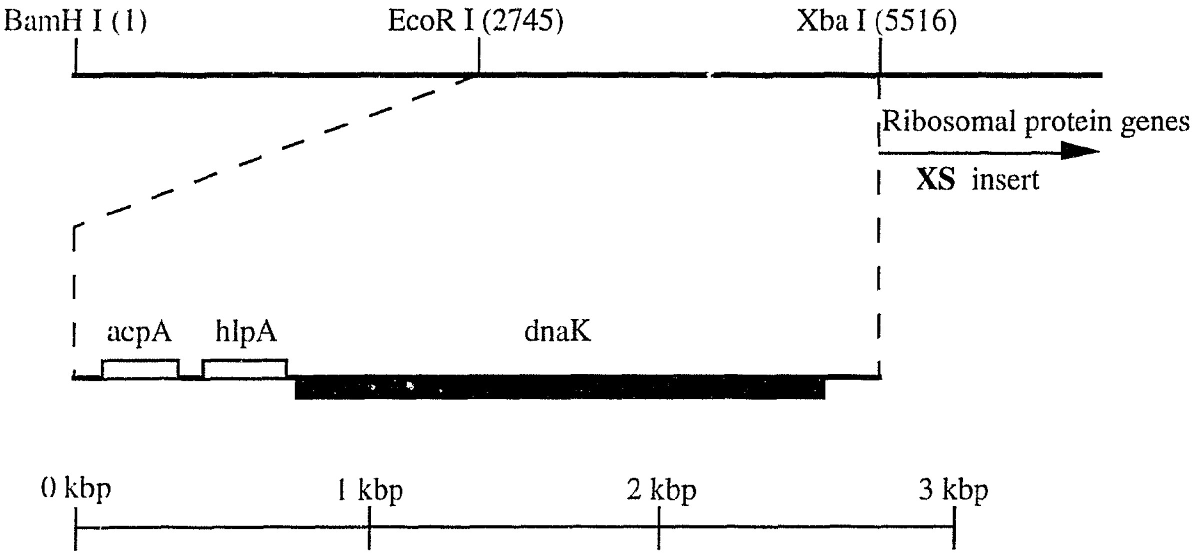

BamH I (1)
EcoR I (2745)
Xba I (5516)
Ribosomal protein genes
XS insert
acpA
hlpA
dnaK
0 kbp
1 kbp
2 kbp
3 kbp

Figure 4-1

Figure 4-2. The DNA sequence and deduced protein sequences of *acpA*, *hlpA* and *dnaK* genes. DNA sequences in upper case are coding sequences, lower case letters denote intergenic sequences. Deduced amino acid sequences are below the DNA sequences, with the name of the deduced protein above the start codon of each gene and highlighted. The restriction sites at the ends of the 2.8 kbp fragment are underlined. The intergenic sequence between the *hlpA* and *dnaK* genes is shown at the end of *hlpA* gene.

```
EcoRI                                              AcpA
gaattcatttattgtttattgaaataaatataaattattcattaaatttt ATG AAC GAG CAA GAA
                                                   Met asn glu gln glu

ATA TTT GAA AAA GTA CAA ACT ATT ATT TCT GAA CAA TTA GGT GTT GAT AAA AGT
ile phe glu lys val gln thr ile ile ser glu gln leu gly val asp lys ser

CAA GTT ACT AAA GAC GCT AAT TTT GCT AAT GAT TTG GGA GCT GAT TCT TTA GAC
gln val thr lys asp ala asn phe ala asn asp leu gly ala asp ser leu asp

ACT GTT GAG TTA GTA ATG GCA ATT GAA GAA GCT TTT AAT ATT GAA ATT CCT GAT
thr val glu leu val met ala ile glu glu ala phe asn ile glu ile pro asp

GAT GCA GCT GAA CAA ATT TCG AAT TTA CAA CAG GCT GTG GAC TTT ATC AGT CAA
asp ala ala glu gln ile ser asn leu gln gln ala val asp phe ile ser gln

AAA GTT GCC GCT TAA tcaatattgatagttatttgtaaaattaaatattttaagtaaagttgtg
lys val ala ala OCH
                                                        HlpA
attttacttaaatattattcgtttctaaactgtaaaatttttaatcatgttat   ATG AAT AAA TCC
                                                        Met asn lys ser

CAG TTA ATT TCT AAG ATA GCA TAT TAC ACA AAA TAT TCG AAA ACT GAT ATT GAA
gln leu ile ser lys ile ala tyr tyr thr lys tyr ser lys thr asp ile glu

AAA ATT ATT ACT AGT ATG CTC GAA ATT ATT GTA GAT ACA GTT GCA ACC GGT GAA
lys ile ile thr ser met leu glu ile ile val asp thr val ala thr gly glu

AAA GTT ACT TTA GTT GGT TTT GGA TCT TTT GAA GCC CGT GAA CGG AAA GCT CGA
lys val thr leu val gly phe gly ser phe glu ala arg glu arg lys ala arg

GAA GGA CGA AAT CCT AGA ACA GGT GAA AAG CTA TTT TTA CCA GCT TCA AGA ATA
glu gly arg asn pro arg thr gly glu lys leu phe leu pro ala ser arg ile

CCA ACT TTT TCC GTA GGA AAT TTT TTT CGA AAC AAA GTT AAT AAA ACA TTC TAA
pro thr phe ser val gly asn phe phe arg asn lys val asn lys thr phe OCH

ttaatttatttta

Xba I
tctagaaaattgtaaaatatgcaactaacgactatttcataataatatgaataaagaaaaaaaaaataac

ctttagatatctgtaaatgttatttaaatgctataacaacgtcggttgtacctactttttaaatacacacc

                                                        DnaK
aacaagataagatcttataaatactaaaaatgatttaaaagtttacattaattcatATG GGA AAA GTA
                                                        Met gly lys val

GTT GGT ATT GAT TTA GGT ACT ACA AAT TCA GTA GTG GCA GTG ATG GAA GGT GGA
val gly ile asp leu gly thr thr asn ser val val ala val met glu gly gly
```

Figure 4-2 continues

```
AAA CCT GCG GTA ATT CAA AAT GCT GAA GGA TTT AGA ACT ACA CCT TCA GTT GTC
lys pro ala val ile gln asn ala glu gly phe arg thr thr pro ser val val

GCA TAC ACA AAA ACA GGA GAT AGA TTG GTT GGT CAA ATA GCC AAA AGA CAA GCT
ala tyr thr lys thr gly asp arg leu val gly gln ile ala lys arg gln ala

GTT ATT AAT CCC GAT AAT ACA TTT TAT TCG GTG AAA AGA TTC ATA GGC CGG CGT
val ile asn pro asp asn thr phe tyr ser val lys arg phe ile gly arg arg

TCA GAA GAA GTG TCA GAA GAA TTA AAA CAA GTA TCT TAT ATT GTA AAA ACA GAT
ser glu glu val ser glu glu leu lys gln val ser tyr ile val lys thr asp

AGT AAT GGA AAT ATT AAA TTA GAT TGT CCA TCA TTA AAA AAG GAA TTT GCT TCA
ser asn gly asn ile lys leu asp cys pro ser leu lys lys glu phe ala ser

GAA GAA ATA TCT GCT GAA GTA TTA CGA AAA TTA GTA GAT GAT GCT AGT AAA TAC
glu glu ile ser ala glu val leu arg lys leu val asp asp ala ser lys tyr

CTT GGA GAA TCT GTA AAA CAA GCG GTA ATT ACA GTT CCA GCT TAC TTT AAT GAT
leu gly glu ser val lys gln ala val ile thr val pro ala tyr phe asn asp

TCG CAA AGG CAA GCA ACT AAA GAT GCA GGT CGA ATT GCA GGT TTA GAA GTA CTA
ser gln arg gln ala thr lys asp ala gly arg ile ala gly leu glu val leu

AGA ATT ATA AAT GAG CCT ACA GCA GCT TCT TTA GCT TAC GGT TTG GAT AAA AAA
arg ile ile asn glu pro thr ala ala ser leu ala tyr gly leu asp lys lys

AAT AAT GAA ACC ATA TTA GTG TTT GAT TTA GGT GGA GGA ACA TTT GAT GTT TCA
asn asn glu thr ile leu val phe asp leu gly gly gly thr phe asp val ser

GTA TTA GAG GTT GGA GAT GGA GTT TTC GAA GTT TTG TCA ACA TCA GGT GAC ACT
val leu glu val gly asp gly val phe glu val leu ser thr ser gly asp thr

CAT TTG GGT GGA GAT GAT TTT GAT GAT AAG ATT GTA CAG TGG TTG TTA AAA GAG
his leu gly gly asp asp phe asp asp lys ile val gln trp leu leu lys glu

TTT GAA ACA GAA CAC AGT ATA AAC TTA AAA TCC GAT CGG CAA GCT TTA CAA AGG
phe glu thr glu his ser ile asn leu lys ser asp arg gln ala leu gln arg

TTA ACA GAA GCG TCG GAA AAA GCA AAA ATT GAA TTA TCC AAC TTA AGT CAA ACT
leu thr glu ala ser glu lys ala lys ile glu leu ser asn leu ser gln thr

GAG ATT AAT TTA CCT TTT TTA ACA GCG ACT GAA ACT GGG CCA AAA CAT TTA GAA
glu ile asn leu pro phe leu thr ala thr glu thr gly pro lys his leu glu

CGT TCA ATA ACT AGA GCA AAA TTT GAG GAA TTA TGT TCT GAT TTA ATT AAT CGA
arg ser ile thr arg ala lys phe glu glu leu cys ser asp leu ile asn arg

GTA AAA ATA CCA GTT GAA AAT GCT TTA AAA GAT GCA AAA TTA GAC TCA AGT AAA
val lys ile pro val glu asn ala leu lys asp ala lys leu asp ser ser lys
```

Figure 4-2 continues

```
ATT GAT GAA GTA GTA TTA GTT GGT GGA TCA ACT CGT ATA CCG GCT ATT CAA GAA
ile asp glu val val leu val gly gly ser thr arg ile pro ala ile gln glu

CTA GTA AAA AGA ATA TTA AAT AAA ACA CCA AAT CAA ACA GTT AAT CCT GAT GAA
leu val lys arg ile leu asn lys thr pro asn gln thr val asn pro asp glu

GTA GTA GCA ATT GGA GCT GCT GTG CAA GCT GGT GTT TTA GCA GGT GAA GTA AAA
val val ala ile gly ala ala val gln ala gly val leu ala gly glu val lys

GAT ATA TTA TTA CTA GAT GTA ACA CCG TTA TCA TTA GGT GTT GAA ACA TTA GGT
asp ile leu leu leu asp val thr pro leu ser leu gly val glu thr leu gly

GGC GTG ACA ACA AGA ATA ATA CCA AGA AAT ACA ACA ATA CCT ACT AAA AAA TCT
gly val thr thr arg ile ile pro arg asn thr thr ile pro thr lys lys ser

GAA GTA TTT TCA ACT GCC GTA GAT AAT CAA CCT AAT GTT GAG ATA CAT GTA TTA
glu val phe ser thr ala val asp asn gln pro asn val glu ile his val leu

CAA GGT GAA CGT GAA TTT GCG AAA GAT AAC AAA AGC TTA GGT ACA TTT CGA TTA
gln gly glu arg glu phe ala lys asp asn lys ser leu gly thr phe arg leu

GAT GGT ATA TTG CCA GCT CCA AGA GGA GTA CCA CAA ATT GAG GTA ACT TTT GAT
asp gly ile leu pro ala pro arg gly val pro gln ile glu val thr phe asp

ATA GAT GCA AAT GGA ATA TTA TCT GTA ACT GCA AAA GAT AAA GGT ACA GGT AAA
ile asp ala asn gly ile leu ser val thr ala lys asp lys gly thr gly lys

GAG CAA TCA ATT ACA ATA ACA GGA GCT TCA ACT TTA CCA TCA GAT GAA GTA GAA
glu gln ser ile thr ile thr gly ala ser thr leu pro ser asp glu val glu

CGT ATG GTA AAT GAA GCA CAA AAT AGT GCT AAG GAA GAT AAA GAA AAG AGA GAT
arg met val asn glu ala gln asn ser ala lys glu asp lys glu lys arg asp

AAA ATT GAT CTA AAG AAT CAA AGT GAT TCT TTA TGC TAT CAA TCA GAA AAA CAA
lys ile asp leu lys asn gln ser asp ser leu cys tyr gln ser glu lys gln

CTC AAA GAA TTA GAA GGA AAA ATA GAC GAT ACA AAT AAA AAT AAA ATT AGT TCT
leu lys glu leu glu gly lys ile asp asp thr asn lys asn lys ile ser ser

ATG ATT TCT GAA TTA CGT AAT GCA ATT AAT AAT GAA AAT TAC GAC GAA ATG AGA
met ile ser glu leu arg asn ala ile asn asn glu asn tyr asp glu met arg

GAT CTA AAT TCA AAA CTA CAA ACA GCT TTA ATG GAT TTA GGC AAA AGC GTT TAT
asp leu asn ser lys leu gln thr ala leu met asp leu gly lys ser val tyr

GAA AAA ACA AGC AAA GAA CAA ACA AGT ACA AGT TCA CCT ACA AAC TCA AAC GAT
glu lys thr ser lys glu gln thr ser thr ser ser pro thr asn ser asn asp

AGC GTA ATA GAT GCA GAT TTT TCA GAA ACA AAA TGA
ser val ile asp ala asp phe ser glu thr lys OPA
```

Figure 4-2

Protein sequence comparisons revealed significant sequence similarity between AcpA and several known acyl carrier proteins (Fig. 4-3). Amino acid sequence identities between the *Guillardia theta* plastid AcpA and the others are 56% for *E. coli* (Vanaman et al, 1968), 59% for *Anabaena variabilis* (Froehlich et al, 1990), 44% for spinach Acp-I (Kao and Ohlrogge, 1984), 36% for spinach Acp-II (Schmid and Ohlrogge, 1990), 43% for *Arabidopsis thaliana* (Post-Beittenmiller et al, 1989) and *Brassica campestris* (Rose et al, 1987) Acps. A serine residue, which is conserved in all the acyl carrier proteins and functions as the attachment site for the prosthetic group phosphopantotheine (Jaworski et al, 1989), is also present in the AcpA sequence. These findings clearly identified *acpA* as an acyl carrier protein gene. Acyl carrier protein is a key cofactor in the synthesis and metabolism of fatty acids (Schmid and Ohlrogge, 1990). The finding of the *acpA* gene in the plastid genome of *Guillardia theta* identifies lipid biosynthesis as a metabolic pathway involving plastid-encoded proteins.

C. Identification of the histone-like protein gene (*hlpA*)

The *hlpA* gene predicts a polypeptide (HlpA) of 93 amino acid residues long, which contains 9 acidic, 17 basic, and 31 hydrophobic amino acid residues. The HlpA is a small and very basic protein, having a calculated molecular mass of 10.6-kDa and a calculated pI of 10.48. Protein sequence comparisons revealed significant sequence similarity between HlpA and several known histone-like proteins (Fig. 4-4). Amino acid sequence identities between the *Guillardia theta* plastid HlpA and the others are 25-53%. This sequence identity increases to 39-76% if the comparison is limited to the region between residues 38 and 79 (Table 4-1), which corresponds to the DNA-binding long arm of histone-like proteins (Tanaka et al, 1984, Drlica and Rouviere, 1987). These findings clearly identified *hlpA* as a histone-like protein gene.

Figure 4-3. Comparison of predicted AcpA protein sequence with known acyl carrier protein sequences. The deduced amino acid sequence of the *Guillardia theta* plastid AcpA is aligned with sequences of known acyl carrier proteins of *E. coli* (E. c.), *Anabaena variabilis* (A. v.), Spinach ACP-I (SA-I), Spinach ACP-II (SA-II), *Arabidopsis thaliana* (A. t), and *Brassica campestris* (B. c). Dots represent residues that are identical to the corresponding ones in the *Guillardia theta* plastid AcpA sequence. Dashes represent computer-generated gaps that maximize the alignment. The universally conserved serine residue is highlighted by a star.

```
                  10        20         30        40        50
AcpA      MNEQEIFEKVQTIISEQLGVDKSQ-VTKDANFANDLGADSLDTVELVMAIEEAFN
E. coli      ST.E.R.KK..G.....KQEE-..DN.S.VE...............L..E.D
A. v.      SQS.T....KK.VI...S.ENPDT..PE.S.....Q............L..E.D
SP Acp-I   AKK.TID..CD.VK.K.ALGADVV..A.SE.S-K..........I..NL..E.G
SP Acp-II AAKP.MVT..SD.VKS..ALAEDAK..GETK.S-EI.........I..KL..E.G
A. t.     AAK..TI...SA.VKK..SLTPDKK.VAETK..-...........I..GL..E..
B. c.     AAKP.TV...SK.VKK..SLKDD.K.VAETK..-...........I..GL..E.D
                                                 *

              60        70        80
AcpA      IEIPDDAAEQISNLQQAVDFISQKVAA
E. coli   T....EE..K.TTV.A.I.Y.NGHQ.
A. v.     .....E...K.TTV.A.....N
SP Acp-I  .NVDE.K.QD..TI...A.V.ESLLEKK
SP Acp-II VTVEEEN.QT.TTI.E.A.M.EALQQNK
A. t.     .QMAEEK.QK.ATVE..AEL.EELINEKK
B. c.     ..MAEEK.QK.ATVEE.AEL.EEL.QLKK
```

Figure 4-3

Figure 4-4. Sequence comparison of *Guillardia theta* plastid HlpA protein with bacterial histone-like proteins. The amino acid sequence of HlpA, deduced from the *Guillardia theta* plastid *hlpA* gene, is aligned with sequences of bacterial histone-like proteins: HAn of *Anabaena* 7120; HCp of *Clostridium pasteurianum*; HBs of *Bacillus stearothermophilus*; Huα, IHF-α and IHf-β of *Escherichia coli*; Tf1 of *Bacillus subtilis* bacteriophage SP01; HRm of *Rhizobium meliloti*; and HTa of *Thermoplasma acidophilum*. Dots represent residues that are identical to the corresponding ones in the *Guillardia theta* plastid HlpA sequence. Dashes represent computer-generated gaps that maximize the alignment.

```
                    10         20        30        40        50        60
HlpA        MNKSQLISKIA-YYTKYSKTDIEKIITSMLEIIVDTVATGEKVTLVGFGSFEARERKAREG
HAn         ...GE.VDAV..EKASVT.KQADAVL.AA..T.IEA.SR.D...........S........
HCp         ...AE..TSM..EKS.LT.K.A.LALKALI.SVEEALEK....Q.....T..T...A....
HBs         ...TE..NAV..ETSGL..K.AT.AVDAVFDS.TEALRK.D..Q.I...N..V...A..K.
HUb         .......D....AGADI..AAAGRALDAIIASVTESLKE.DD.A.....T.AVK..A..T.
HUa         ...T...DV...EKAEL...QAKAALE.T.AA.TESLKE.DA.Q.....T.KVNH.AE.T.
IHFb        .T..E..ERL.TQQSHIPAKTV.DAVKE...HMAS.L.Q..RIEIR.....SLHY.AP.T.
IHFa      MALT.AEMSEYLF.DKLGL..R.AKELVELFF.E.RRALEN..Q.K.S...N.DL.DKNQ.P.
Tf1         ...TE..KA...QD.ELTQVSVS.MLA.FEK.TTE...K.D..Q.T..LNIKPVA.Q..K.
HRm         ...NE.VAAV..DKAGL..A.ASSAVDAVF.T.QGELKN.GDIR.....N.SVSREASKGR
HTa        MVGI.E.SKEV..KKANTTQKVARTV.K.F.DE..SEANG.Q.IN.A...I..R.TQFP.KA


                  70        80        90
HlpA      RNPRTGEKLFLPASRIPTFSVGNFFRNKVNKTF
HAn       ...K.N..MEI..T.V.A..A.KL..E..APPKA
HCp       .....K VINI..TTV.V.KA.KE.KD....
HBs       ...Q...EMEI...KV.A.KP.KALKDA.K
HUb       ...Q..KEITIA.AKV.S.RA.KALKDA..
HUa       ...Q..KEIKIA.ANV.A.VS.KALKDA.K
IHFb      ...K..D.VE.EGKYV.H.KP.KEL.DRNIYG
IHFa      ...K...DIPIT.R.VV..RP.QKLKSR.ENASPKDE
Tf1       F..Q.Q.A.EIAP.VGVSVKP.ESLKKAAEGLKYEDFAK
HRm       ...S..AEVDI..RNV.K.TA.KGLKDA..
HTa       ...Q.KKVIEV.SKKKFV.RASSKIKYQQ
```

Figure 4-4

	HAn	HCp	HBs	HU-β	HU-α	IHF-β	IHF-α	Tf1	HRm	HTa
HlpA	53	47	41	37	37	35	29	31	30	25
HlpA-(38-79)	(76)	(68)	(63)	(51)	(46)	(46)	(51)	(39)	(39)	(39)

Table 4-1 The percentage of identical amino acid residues between plastid HlpA and other histone-like proteins. Numbers are the percentage of identical amino acid residues between the plastid histone-like protein and that of another organism. Numbers in parenthesis are the percentage of identical amino acid residues when the comparison is limited to sequence between residues 38 and 79.

D. Identification of the Hsp70-like heat shock protein gene (*dnaK*)

DNA sequence of the *dnaK* gene and the protein sequence predicted from it are shown in Fig. 4-2. The predicted protein (DnaK) is 627 amino acid residues long, containing 98 acidic, 78 basic, and 184 hydrophobic amino acid residues. The calculated molecular mass of DnaK is 68.5-kDa. Protein sequence comparisons revealed significant sequence similarity between DnaK and several known Hsp70-like heat shock proteins (Fig. 4-5). Amino acid sequence identities between the *Guillardia theta* plastid DnaK and the others range from 43 to 54%. The sequence similarity is slightly higher when compared to its prokaryotic counterpart (50 to 54%) than to its eukaryotic counterpart (43 to 46%) (Table 4-2). Interestingly, the extensive similarities among these proteins are observed only at the N-terminal 80% of the sequence (52 -63% sequence identity within the N-terminal 500 amino acid residues), while the C-terminal 20% of the sequence (beyond residue 500) shows little conservation (Fig. 4-5). Similarly, when the sequence comparison is limited to the N-terminal 500 amino acid residues, the DnaK protein shows higher similarity to its prokaryotic counterparts (58 to 63%) than to its eukaryotic ones (52 to 56%).

Figure 4-5. Comparison of the DnaK protein with Hsp70 family proteins. Amino acid sequence of the plastid DnaK protein (P. DnaK), deduced from the *Guillardia theta* plastid *dnaK* gene, is aligned with sequences of seven biologically distinct members of the Hsp70 family: one eubacterial member (E. DnaK for *Escherichia coli* DnaK), two mitochondrial members (SscI of yeast and Mtp70 of *Trypanosoma cruzi*, both nucleus-encoded), and four members that function in the cytoplasm of eukaryotic cells (Tryp for *Trypanosoma brucei* Hsp70, yeast SsaI, maize and human Hsp70). Dots represent residues that are identical to the corresponding ones in the *Guillardia theta* plastid DnaK sequence. Dashes represent computer-generated gaps that maximize the alignment.

```
                                            10        20        30
P. DnaK                             MGKVVGIDLGTTNSVVAVMEGGKPAVIQNAEGFRTT
E. DnaK                             ...II.........C..I.D.TT.R.LE....D...
Ssc1    MLAAKNILNRSSLSSSFRIATRLQSTKVQ.S.I.........A..I...KV.KI.E....S...
Mtp70      MFARRLRGAGSLAAASLARWQSSKVT.D.I.......Y.C......D..R.LE.T....A.
Tryp.                            MTYE.-AI.......Y.C.G.WQNERVEI.A.DQ.N...
Yeast                               .S.A........Y.C..HFANDRVDI.A.DQ N ..
Maize                           MAKSE.PAI.......Y.C.GLWQHDRVEI.A.DQ.N...
Human                             MAKAAA........Y.C.G.FQH..VEI.A.DQ.N...

          40        50        60        70          80        90
P. DnaK PSVVAYTKTGDRLVGQIAKRQAVINPDNTFYSVKRFIGRR--SEEVSEELKQVSYIVKTDSNGN
E. DnaK ..II...QD.ET....P......T..Q..LFAI..L....FQD...QRDVSIMPFKIIAAD..D
Ssc1    .....F..E.E....IP......V..E..LFAT..L....FEDA..QRDI...P.KIVKH...D
Mtp70   .....F-.GQEK...LA......T..QS..FA...L....FEDSNIQHDI.N.P.KIGRS...D
Tryp.   ..Y..F.-DSE..I.DA..N.VAM..T..VFDA..L...KFSDSV.QSDM.HWPFK.V.KGDDK
Yeast   ..F..F.-DTE..I.DA..N..AM..S..VFDA..L...NFNDP..QADM.HFPFKLI-.VD.K
Maize   ..Y.GF.-DTE..I.DA..N.VAM..T..VFDA..L....FS.PA.QSSM.LWPSRHL-GLGDK
Human   ..Y..F.-DTE..I.DA..N.VAL..Q..VFDA..L...KFGDPV.QSDM.HWPFQ.IN-DGDK

       100         110       120       130       140       150       160
P. DnaK IKL--DCPSLKKEFASEEISAEVLRKLVDDASKYLGESVKQAVITVPAYFNDSQRQATKDAGRI
E. DnaK AWV--EVKG--QKM.PPQ......K.MKKT.ED....P.TE...........A...........
Ssc1    AWV--EARG--QTYSPAQ.GGF..N.MKET.EA...KP..N..V..................Q.
Mtp70   AWVQ-.ANG--.QYSPSQVG.F..E.MKET.ENF..RK.SN..V.C.....GP.........T.
Tryp.   PVIQVQFRGET.T.NP....SM..L.MKEV.ES...KQ.AK..V..................T.
Yeast   PQIQVEFKGET.N.TP.Q..SM..G.MKET.ES...AK.ND..V..................T.
Maize   PMIVFNYKGEE.Q..A....SM..I.MKEI.EA...STI.N..V..................V.
Human   P.VQVSYKGET.A.YP....SM..T.MKEI.EA...YP.TN.....................V.

               170       180           190       200           210
P. DnaK AGLEVLRIINEPTAASLAYGLDKK-N---NETILVFDLGGGTFDVSVLEV----GDGVFEVLST
E. DnaK .....K.........A.......GTG---.R..A.Y........I.II.IDEVD.EKT....A.
Ssc1    V..N...VV......A.....E.S-D---SKVVA..........I.I.DI----DN.....K..
Mtp70   ...N.I.VV.G....A.......T-K---DSM.A.Y........I....I----AG.....KA.
Tryp.   ...............AI......ADEGK-ERNV.I..........TL.TI----DG.I...KA.
Yeast   ...N...........AI......--KGK-E.HV.I...........L.FI----E..I...KA.
Maize   ...N.M.........AI.......ATSSGEKNV.I...........L.TI----EE.I...KA.
Human   ...N...........AI.....RTG--KGERNV.I...........I.TI----D..I...KA.

         220       230         240      250       260       270
P. DnaK SGDTHLGGDDFDDKIVQWLLKEFETEHS-INLKSDRQALQRLTEASEKAKIELSNLSQTEINLP
E. DnaK N.......E...SRLINY.VE..KKDQG-.D.RN.PL.M...K..A........SAQ..DV...
Ssc1    N.......E...IYLLREIVSR.K..TG-.D.EN..M.I..IR..A........STVS......
Mtp70   N.......E...LCLSDYI.T..KKSTG-.D.SNE.M....IR..A....C...TTME..V...
Tryp.   N.......E...NRL.AHFTE..KRKNKGKD.S.NLR..R..RT.C.R..RT..SAA.AT.EID
Yeast   A.......E...NRL.NHFIQ..KRKNK-KD.STNQR..R..RT.C.R..RT..SSA..SVEID
Maize   A.......E...NRM.NHFVQ..KRKNK-KDISGNPR..R..RT.C.R..RT..STA..T.EID
Human   A.......E...NRL.NHFVE..KRK.K-KDISQNKR.VR..RT.C.RFEGIDFYT.I.RARFE
```

Figure 4-5 continues

```
      280       290       300       310       320       330       340
P. DnaK FLTATETGPKHLERSITRAKFEELCSDLINRVKIPVENALKDAKLDSSKIDEVVLVGGSTRIPA
E. DnaK YI..DA.....MNIKV....L.S.VE..V..SIE.LKV..Q..G.SV.D..D.I....Q..M.M
Ssc1    .I..DAS....INMKFS..Q..T.TAP.VK.TVD..KK.....G.ST.D.S..L....MS.M.K
Mtp70   .I..NQD.AQ.VQMTVS.S...S.AEK.VQ.SLG.CKQCI...AV.LKE.S.......M..M.K
Tryp.   A.FENI----DFQAT....R.....G..FRGTLQ...RV.Q...M.KRAVHD...........K
Yeast   S.FEGI----DFYT.....R.....A..FRSTLD...KV.R.....K.QV..I..........K
Maize   S.FEGI----DFTPRSS..R....NM..FRKCME...KC.R...M.K.SVHD...........K
Human   E.AKRT----LSSSTQASLEIDS.....FRSTLE...K..R.....KAQ.HDL..........K

            350        360       370       380            390       400
P. DnaK IQELVKRIL-NKTPNQTVNPDEVVAIGAAVQAGVLAGEVK-----DILLLDVTPLSLGVETLGG
E. DnaK V.KK.AEFF-G.E.RKD.....A........G...T.D..-----.V...........I..M..
Ssc1    VV.T..SLF-G.D.SKA.....A........GA..S...T-----.V...........I.....
Mtp70   VI.A..QFF-GRD.FRG.....A..L.G.TLG...RRD..-----GLV................
Tryp.   VMQ..SDFFGG.EL.KSI....A.XY......FI.T.GKSKQT-EGL.....A..T..I..A..
Yeast   V.K..TDYFNG.E..RSI....A..Y......AI.T.DESSKT-Q.L.....A.....I..A..
Maize   V.Q.-QDFFNG.ELCKSI....A..Y......AI.S..GNERS--.L...........L..A..
Human   V.K.LQDFFNGRDL.KSI....A.GY......AI.M.DKSENV-Q.L.....A.....L..A..

              410        420       430       440       450       460
P. DnaK VTTRIIPRNTTIPTKKSE-VFSTAVDNQPNVEIHVLQGEREFAKDNKSLGTFRLDGILPAPRGV
E. DnaK .M.TL.AK.......H.Q-.....E...SA.T........KR.A......Q.N....N.....M
Ssc1    .F..L............Q-I....AAG.TS...R.F.....LVR...LI.N.T.A..P...K..
Mtp70   .F..M..K.........QTF....AF..TQ.G.K.F.....M.A..QMM.Q.D.V..P......
Tryp.   .M.AL.K..........Q-T...YS....G.H.Q.FE...TMT..CHL....D.S..P......
Yeast   .M.KL....S..S...F.-I...YA....G.L.Q.FE...AKT...NL..K.E.S..P......
Maize   .M.VL...........EQ-....YS....G.L.Q.YE...ART...NL..K.E.S..P......
Human   .M.AL.K..S.....QTQ-I.T.YS....G.L.Q.YE...AMT...NL..R.E.S..P...-..

           470       480       490       500        510       520
P. DnaK PQIEVTFDIDANGILSVTAKDKGTGKEQSITITGAST-LPSDEVERMVNEAQNSAKEDKEKRDK
E. DnaK ...........D...H.S....NS....K...KAS.G-.NE..IQK..RD.EAN.EA.RKFEEL
Ssc1    ...........D..IN.S.R..A.N.DS...VA.S.G-.SEN.I.Q...D.EKFKSQ.EARKQA
Mtp70   .........EP...CH......A...T.N....ASGG-.SKEQI...IRDSESH.ES.RL..EL
Tryp.   ........L........S.EE.....RNQ.V..NDKGR.SKADI....SD.AKYEA...AHVXX
Yeast   ........V.S....N.S.VE.....SNK....NDKGR.SKEDI.K..A..EKFKE..EKESQR
Maize   ...T......V.N..N.S.E..T..QKNK....NDKGR.SKE.I.K..Q..EKYKA..E.VKK.
Human   ...............N...T..S...ANK....NDKGR.SKE.I....Q..EKYKA..EVQ.ER

        530       540         550       560       570          580
P. DnaK IDLKNQSDSLCYQSEKQLKE--LEGLIDDTNKNKISSMISELRNAINNENY---DEMRDLNSKV
E. DnaK VQTR..G.H.LHSTR..VE.--AGDKLPADD.TA.E.ALTA.ET.LKG.DKA---AIEAKMQEL
Ssc1    .ETA.KA.Q.ANDT.NS...--F..KV.KAEAQ.VRDQ.TS.KELVARVQGG--E.VNA--EEL
Mtp70   VEVR.NAETQANTA.R..T.--WK-YVS.AE.ENVRTLLRAC.KSME.P.V.TK..LSAATD.L
Tryp.   ..A..GLENYAFSMKNTINDPNVA.KL..AD..AVTTAVE.ALRWL.DNQEASLE.YNHRQKEL
Yeast   .AS...LE.IA.SLKNTIS.--AGDKLEQAD.DTVTKKAE.TISWLDSNTTASKE.FD.KLKEL
Maize   V.A..ALENYA.NMRNTI.DDKIASKLPAED.K..EDAVDGAISWLDSNQLAEVE.FE.KMKEL
Human   VSA..ALE.YAFNMKSAVEDEG.K.K.SEAD.K.VLDKCQ.VISWLDANTLAEK..FEHKRKEL
```

Figure 4-5 continues

```
         590       600                 610       620
P. DnaK QTALMDLGKSVYEKTSKQQTSTS----------SPTNSNDSVIDADFSETK
E. DnaK AQVSQK.MEIAQQQHAQ...AGAD-------ASANNAKD.D.V..E.E.V.DKK
Ssc1    K.KTEE.QT.SMKLFEQLYKND.-----------NN.N.NNGNN.ESG...Q
Mtp70   .K.V.EC.RTE.QQAAAGNS.S.---------SGN.D.SQGEQQQQGDQQ.Q
Tryp.   EGVCAPILSKM.QGMGG--GDGPGGMPEGMPGGM.GGMPGGMGGGMGGAAASSGPKVEEVD
Yeast   .DIANPIMSKL.QA-----GGAPGGAAGGAPGGF.GGAPPAPEAEGPTVEEVD
Maize   EGICNPIIAKM.X-------GEGAG--MGAAAGMDEDAPSGGSG.GPKIEEVD
Human   EQVCNPIISGL.QGA----GGPGPG-----GFGAQGPKGG.GSGPTIE.VD
```

Figure 4-5

Table 4-2. Sequence similarity between plastid DnaK and other Hsp70 proteins

	Plastid DnaK	*E. coli* DnaK	Sce1	Mtp70	Tryp	Yeast	Maize
E.coli DnaK	54						
Ssc1	53	56					
Mtp70	50	53	57				
Tryp	43	43	44	42			
Yeast	46	46	46	44	69		
Maize	44	45	45	41	68	70	
Human	44	45	44	42	65	67	67

Numbers shown are percentages of identical amino acid residues between individual pairs of sequences. Calculations are based on the sequence alignment in Figure 4-5.

DISCUSSION

Protein sequence comparisons clearly identified the *acpA* gene product as an acyl carrier protein, which is a key cofactor in the synthesis and metabolism of fatty acids (Schmid and Ohlrogge, 1990). This is the first time an *acpA* gene has been found in an organelle genome, thus demonstrating for the first time that lipid biosynthesis as a metabolic pathway in plastids can involve plastid-encoded proteins. It is noted that the AcpA protein is most similar to the acyl carrier protein of *Anabaena variabilis*, which is consistent with a cyanobacterial origin for the plastid of this organism.

Sequence and other structural features of the plastid HlpA protein clearly resemble those of bacterial histone-like proteins. It is relatively small in size and very basic, with a calculated molecular mass of 10.6-kDa and a pI of 10.48. Sequence identities between the plastid HlpA protein and bacterial histone-like proteins are 25-53%. If the comparison is limited to the region between residues 38 and 79, which corresponds to the DNA-binding long arm of histone-like proteins (Tanaka et al, 1984, Drlica and Rouviere, 1987), the protein sequence of the plastid HlpA protein is 76% identical to that of cyanobacterial histone-like protein HAn and 39-68% identical to those of other bacterial histone-like proteins. This family of proteins was called "histone-like" simply because they are small in size, abundant in cells, and very basic. Their cellular functions are more complex than that of eukaryotic nuclear histones. In *E. coli*, it has shown that histone-like proteins are involved in genome organization, stimulation of transcription, site-specific recombination, and initiation of DNA replication (Flashner and Gralla, 1988, Pettijohn, 1988, Schmid, 1990). Because of the structural and sequence similarities of plastid histone-like protein to bacterial histone-like proteins, it is likely that the plastid HlpA protein is involved in organization of plastid genome as well as other processes such as DNA recombination, DNA replication and transcription. Histone-like proteins have previously been extracted from chloroplasts (Briat et al, 1984) and mitochondria (Yamada et al, 1991), although their sequences and genes (most likely encoded in nucleus) were not studied. This work demonstrated for the first time that a histone-like protein is encoded in a plastid genome. Furthermore, the histone-like protein described here is most similar to the histone-like protein of *Anabaena*, which is consistent with a cyanobacterial origin for the plastid of this organism (Gray and Doolittle, 1982, Gray, 1992).

The *Guillardia theta* plastid DnaK protein is clearly a member of the Hsp70 protein family (Fig. 4-5), which is a group of highly conserved proteins found in eukaryotes as well as in prokaryotic cells. Proteins immunologically related to the *E. coli* DnaK protein have been detected previously in chloroplasts of plants and *Euglena*, although their sequences and genes were not studied (Marshall et al, 1990; Amir et al., 1990). The absence of this gene in the completely sequenced chloroplast genomes of several plants and *Euglena*, however, indicates that it is encoded in the nucleus of these organisms. The *Guillardia theta* plastid DnaK protein is more similar to *E. coli* DnaK and the mitochondrial hsp70 proteins than to the eukaryotic cytoplasmic Hsp70 proteins that form a natural group of their own. This is consistent with the notion that genes for both plastid and mitochondrial Hsp70 proteins originated from eubacteria through endosymbiosis, although the mitochondrial Hsp70 protein genes are now in the nucleus (Craig et al., 1989; Engman et al., 1989). The rapid and dramatic increase in the rate of transcription of heat shock protein genes upon stress is initiated by the binding of activated heat shock factor (HSF) to the promoter region of the heat shock protein gene. The HSF trimer binding site on the DNA usually is composed of at least three 5-bp modules (nGAAn) arranged as contiguous inverted repeats (nGAAnnTTCnnGAAn) (Perisic et al., 1989). Such contiguous inverted repeats have been identified for the *Guillardia theta* plastid *dnaK* gene at a position 190-bp upstream of the initiation codon (within the noncoding region upstream of the plastid *dnaK* gene), suggesting that the *dnaK* gene is likely expressed in *Guillardia theta* plastid. Nothing is known about the function of plastid Hsp70 protein. Its structural resemblance to other Hsp70-like proteins, however, suggests similar cellular functions. It has been suggested that the bacterial DnaK proteins are involved in protein translocation across cellular membranes, most likely by modulating the folding and unfolding of other proteins through protein-protein interactions. The *E. coli* DnaK has been implicated also in DNA replication processes (Engman et al., 1989). The plastid DnaK protein is likely, therefore, to function in plastid protein import as well as other activities such as DNA replication.

The three genes described in the plastid genome of *Guillardia theta* were the first such genes to be found in organellar genomes, and each of them represents a functional class of gene products that had not been described previously. The endosymbiont hypothesis proposes that upon endosymbiosis, the majority of genes from the endosymbiont genome were either lost or transferred to the nucleus. Among the retained genes, most of them are for the organelle's specialized functions (photosynthesis in plastids and aerobic respiration in mitochondria) and for its own gene expression (tRNA, rRNA and ribosomal protein genes, etc.). The complete nucleotide sequences of several chloroplast and mitochondrial genomes have been shown to agree well with this scenario, with only a few genes that do not fall into either of these categories. Notable examples in the chloroplast genomes include *clpP* encoding a subunit of the Clp protease and *accD* encoding a subunit of acetyl-CoA carboxylase. The three genes discovered in the plastid genome of *Guillardia theta*, each representing a unique class of functional genes, do not fall into any of the categories described in the chloroplast genome of higher plants and liverwoıt.

Cryptomonads have been demonstrated to have obtained their plastids through secondary endosymbiosis. Study of the *Guillardia theta* plastid genome demonstrated that it is a typical plastid genome, as judged by its topology (double stranded circular DNA molecule), size (within the 100 to 200 kbp range), and GC content (low GC content). However, study of ribosomal protein operons (described in Chapter III) has shown that the *Guillardia theta* plastid genome encodes substantially more genes than plant chloroplast genomes. The discovery of the three genes described in this Chapter added another dimension, because the plastid genome of *Guillardia theta* is different from the chloroplast genomes of land plants not only in the number of genes, but also in the identity of the genes and functions of their gene products. Such differences may have interesting implications in terms of plastid evolution.

Chapter V. Characterization of three chloroplast genes of unusual structure in *Chlamydomonas* and the discovery of intein and protein splicing

INTRODUCTION

As described in Chapter I, the recently discovered inteins and associated protein splicing represent a new dimension in the flow of genetic information from genes to protein. An intein is defined as a protein sequence that is embedded in-frame in a protein precursor and removed at the protein level through protein splicing. Protein snlicing is a post-translational processing event in which the expression of a single gene results in the production of two proteins. One protein corresponds to the internal sequence precisely excised from the protein precursor, while the other corresponds to the two flanking sequences joined together by a bona fide peptide bond. In parallel with RNA splicing, the unspliced protein is called *precursor protein*, which undergoes *protein splicing* to yield an excised polypeptide called *intein* (internal protein sequence), the N- and C-terminal polypeptides, which are called *N-* and *C-exteins* (external protein sequences), are ligated through peptide bond formation to yield the *spliced protein.*

Up to now, only three different proteins were known to contain inteins. They include the nucleus-encoded VMA1 protein of yeast *Saccharomyces cerevisiae* (Kane et al, 1990) and *Candida tropicalis* (Gu et al, 1993), the RecA protein of the eubacteria *Mycobacterium tuberculosis* (Davis et al, 1992) and *Mycobacterium leprae* (Davis et al, 1994), and the DNA polymerase of archaeon *Thermococcus litoralis* (Perler et al, 1992) and *Pyrococcus* species strain GB-D (Xu et al, 1993). It has been suggested that inteins may be more widespread than had been found, partly because the known inteins have characteristics of mobile genetic elements such as endonuclease activities that facilitates intron/intein homing.

The first indication of a intein is usually when a gene sequence predicts a protein much larger than its homologs in other organisms due to the presence of internal extra sequences. For example, the first intein was recognized in *Saccharomyces cerevisiae*, because its VMA1 gene predicted a protein (119-kDa) much larger than a conventional VMA1 protein (69-kDa), and because a large internal stretch of sequence (454 residues) in the predicted VMA1 protein was not similar to sequences of conventional VMA1 proteins or any other known proteins. In another word, the presence of intein can be suspected when a gene contains large translated insertion sequences.

Two *Chlamydomonas* chloroplast genes, *clpP* and *rps3*, contain such large translated insertion sequences (Huang et al., 1994; Liu et al., 1993), suggesting the possible presence of inteins. *Chlamydomonas reinhardtii* is a unicellular green alga that has been studied extensively in terms of cell organelle genetics and molecular biology. It has a single large chloroplast, and its chloroplast genome is a circular double strand DNA molecule of 197 kbp in size. The chloroplast gene content and organization are quite similar to those of land plants, characterized by the presence of a 20-kbp large inverted repeat on which the rRNA genes are located. Most other genes are located in the single copy regions of the genome, including the *clpP* and the *rps3* genes to be described separately below.

The *clpP* gene encodes the catalytic subunit of the ATP-dependent Clp protease. This protease is widespread, if not ubiquitous, among prokaryotes and eukaryotic cell organelles, although only the *E. coli* Clp protease has been characterized extensively in structure and activity. The *E. coli* Clp contains two distinct protein subunits: an 83-kDa regulatory subunit (ClpA) carrying ATPase activity, and a 21-kDa catalytic subunit (ClpP) carrying protease activity. Seven ClpP subunits assemble into a disc-like structure with central cavity, while ClpA forms a disc-like ring of six subunits. The functional protease complex contains two superimposed ClpP rings flanked at one or both ends by a ClpA ring (Kessel et al, 1995). The complex is strikingly similar in its underlying architecture to eukaryotic proteasomes despite the lack of sequence similarity between the two, and it is generally accepted that Clp protease is a prokaryote-type proteasome. Chloroplasts clearly contain a Clp protease similar to that of *E. coli*. In land plants, a *clpP* gene has been found in the chloroplast genome, while a *clpA* homologue (*clpC*) is located in the nuclear genome. The DNA sequence of chloroplast *clpP* gene has been determined in several plants, and they all predict a ClpP protein similar to the *E. coli* ClpP in both size and amino acid sequence.

In *Chlamydomonas*, however, sequence determination of the chloroplast *clpP* gene revealed the presence of large translated insertion sequences (Huang et al., 1994). Unlike *clpP* genes of *E. coli* and other chloroplasts, which encode ClpP protein sequences around 200 residues long, the *Chlamydomonas reinhardtii* chloroplast *clpP* gene predicts a protein sequence of 524 residues, and the *Chlamydomonas eugametos* chloropla‘ *clpP* gene predicts a protein sequence of 1010 residues. Sequence comparisons between *Chlamydomonas* ClpP and non-*Chlamydomonas* ClpP proteins revealed that *Chlamydomonas reinhardtii* ClpP contains a 286-residue insertion sequence (IS1), and that the *Chlamydomonas eugametos* ClpP contains two insertion sequences: IS1 of 318 residues and IS2 of 456 residues (Fig. 5-1). IS1 and IS2 have no obvious sequence similarity to any known proteins, in contrast to their flanking sequences that are very similar to other ClpP proteins. Coding sequences of IS1 and IS2 do not resemble RNA introns and are not excised from the mRNA. Instead, each sequence is most likely translated with its flanking sequences into a single polypeptide. Like the

Figure 5-1. Schematic comparison of protein sequences of *Chlamydomonas reinhardtii* and *Chlamydomonas eugametos* ClpP with those of other organisms. Amino acid sequences of ClpP proteins from *Chlamydomonas reinhardtii*, *Chlamydomonas eugametos* and other organisms are compared schematically. Other organisms include *Marchantia polymorpha, Nicotiana tabacum, Oryza sativa,* and *E. coli.* Black boxes are ClpP protein sequences, open boxes are insertion sequence IS1, and hatched box is insertion sequence IS2. SD stands for sequence domains, IS stands for insertion sequences. The number of amino acid residues i each domain and overali sequence length are indicated.

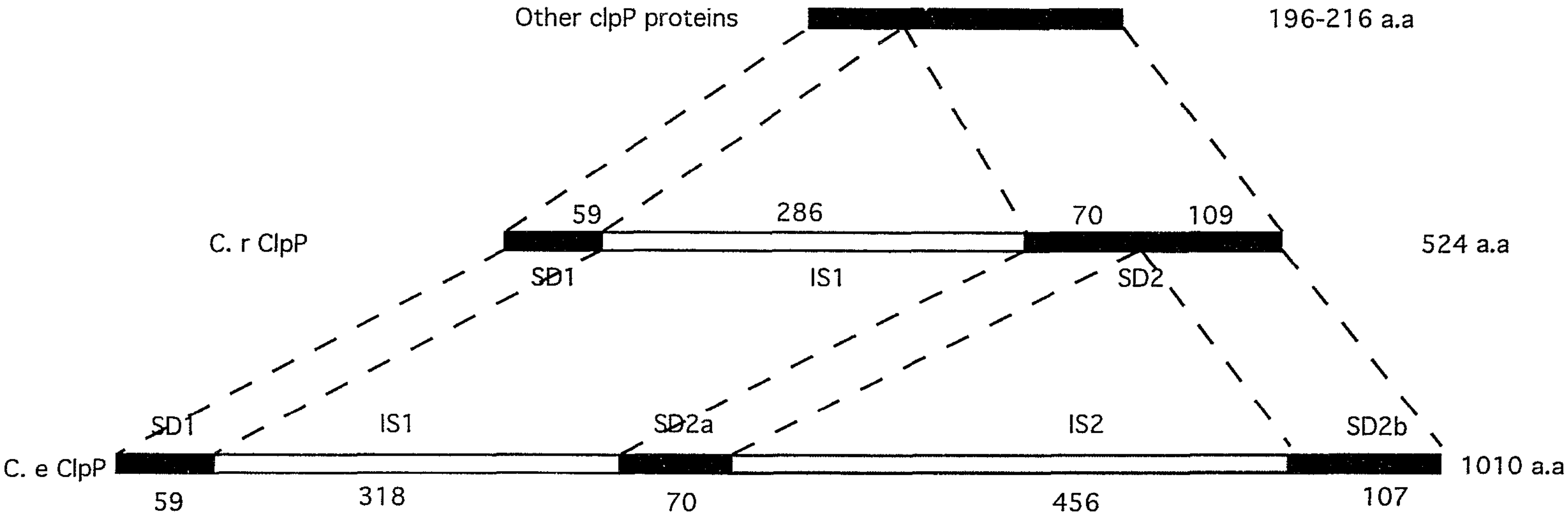

Other clpP proteins
196-216 a.a
C. r ClpP
59
286
70
109
524 a.a
SD1
IS1
SD2
SD1
IS1
SD2a
IS2
SD2b
C. e ClpP
1010 a.a
59
318
70
456
107

Figure 5-1

clpP gene, the *rps3* gene of *Chlamydomonas* chloroplasts also appears to contain a large translated insertion sequence. This gene was initially described in *Chlamydomonas reinhardtii* as an open reading frame (orf712) that lacks a detectable transcript but potentially encodes a polypeptide with sequence similarities to ribosomal protein Rps3 only at its N- and C-termini (Fong et al., 1992). It was subsequently shown to be functionally essential and structurally conserved among various *Chlamydomonas* species, suggesting that it is an unusual *rps3* gene containing a large translated insertion sequence (Liu et al., 1993). A typical *rps3* gene encodes a ribosomal protein (Rps3) located in the 30S subunit of prokaryote-type ribosomes. In *E. coli*, the Rps3 protein sequence is 233 residues long. In various plants, the chloroplast *rps3* gene has been sequenced, in all cases predicting a protein that is similar to *E. coli* Rps3 both in size and in sequence. In *Chlamydomonas reinhardtii* chloroplast, however, the unusual *rps3* gene predicts a protein sequence 712 residues long (Fig. 5-2). When the predicted amino acid sequence was compared to conventional Rps3 sequences of *E. coli* and other chloroplasts, sequence similarity was found at the N- and C-termini, while the middle two-thirds of the *Chlamydomonas* sequence lacks similarity to any known protein sequences.

Based on the above observations, several scenarios have been suggested for the expression of the *Chlamydomonas* chloroplast *clpP* and *rps3* genes. The first and the simplest scenario is that each gene is expressed as a single polypeptide that assembles and functions without further processing. The N- and C-terminal sequence domains may still fold together to form a conventional Rps3 or ClpP protein, with the middle part looping out as spacer sequence of no obvious function. In a second scenario, the precursor protein is processed into three parts (five parts for *Chlamydomonas eugametos* ClpP protein that contains two insertion sequences), with the N- and C- terminal parts assembling and functioning in ribosome or protease as two polypeptides (three polypeptides in the case of ClpP protein in *Chlamydomonas eugametos*). In a third

Figure 5-2. The comparison of Rps3 protein from *E. coli* and *Chlamydomonas reinhardtii* (C.re). The schematic alignment of Rps3 proteins from *E. coli* and *C. reinhardtii*. Black boxes are the four conserved sequence domains of a conventional Rps3 protein, open boxes are sequences of no similarity.

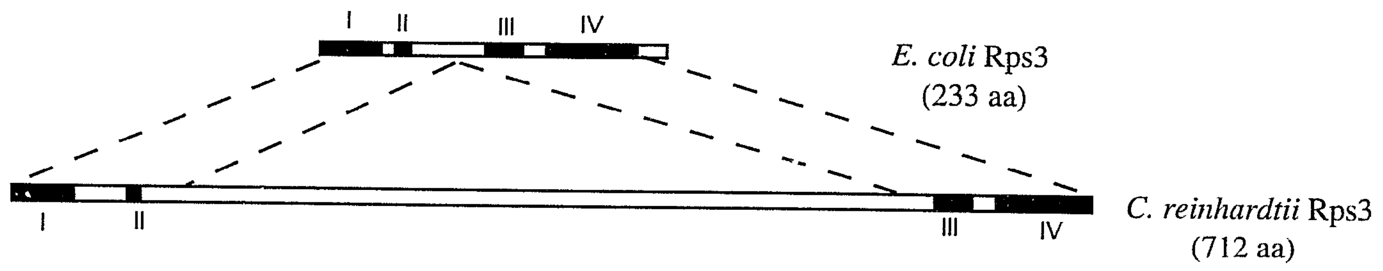

I
II
III
IV
E. coli Rps3
(233 aa)
I
II
III
IV
C. reinhardtii Rps3
(712 aa)

Figure 5-2

scenario, the insertion sequences may be inteins that are removed from the precursor proteins through protein splicing, while the flanking sequence domains are subsequently joined through a peptide bond to form conventional Rps3 or ClpP protein. There have been examples of inteins encoded in eubacterial, archaeal and eukaryotic nuclear genes, but no intein has been described for a cell organelle gene before.

As part of this thesis (to be described in this Chapter), the expression of *clpP* and *rps3* genes in *Chlamydomonas* chloroplast has been investigated. The *Chlamydomonas reinhardtii rps3* gene appears to produce a split protein that is processed into several parts, since four protein products were observed by Western blotting analyses. They included the precursor, the N-terminal part, the S- (insertion sequence) plus the C-terminal part, and the C-terminal part of the protein. All four protein products were assembled into the 30S ribosome subunit and may therefore function in the ribosome. The *Chlamydomonas reinhardtii clpP* gene containing IS1 appears to produce a single polypeptide without further processing, since only the precursor protein was observed both *in vivo* and when the gene was expressed in *E. coli*. The IS2 of *Chlamydomonas eugametos clpP* gene, however, was found to be a degenerate intein whose protein splicing activity can be restored by a single amino acid substitution. This demonstrates for the first time that inteins exist in cell organelles.

RESULTS

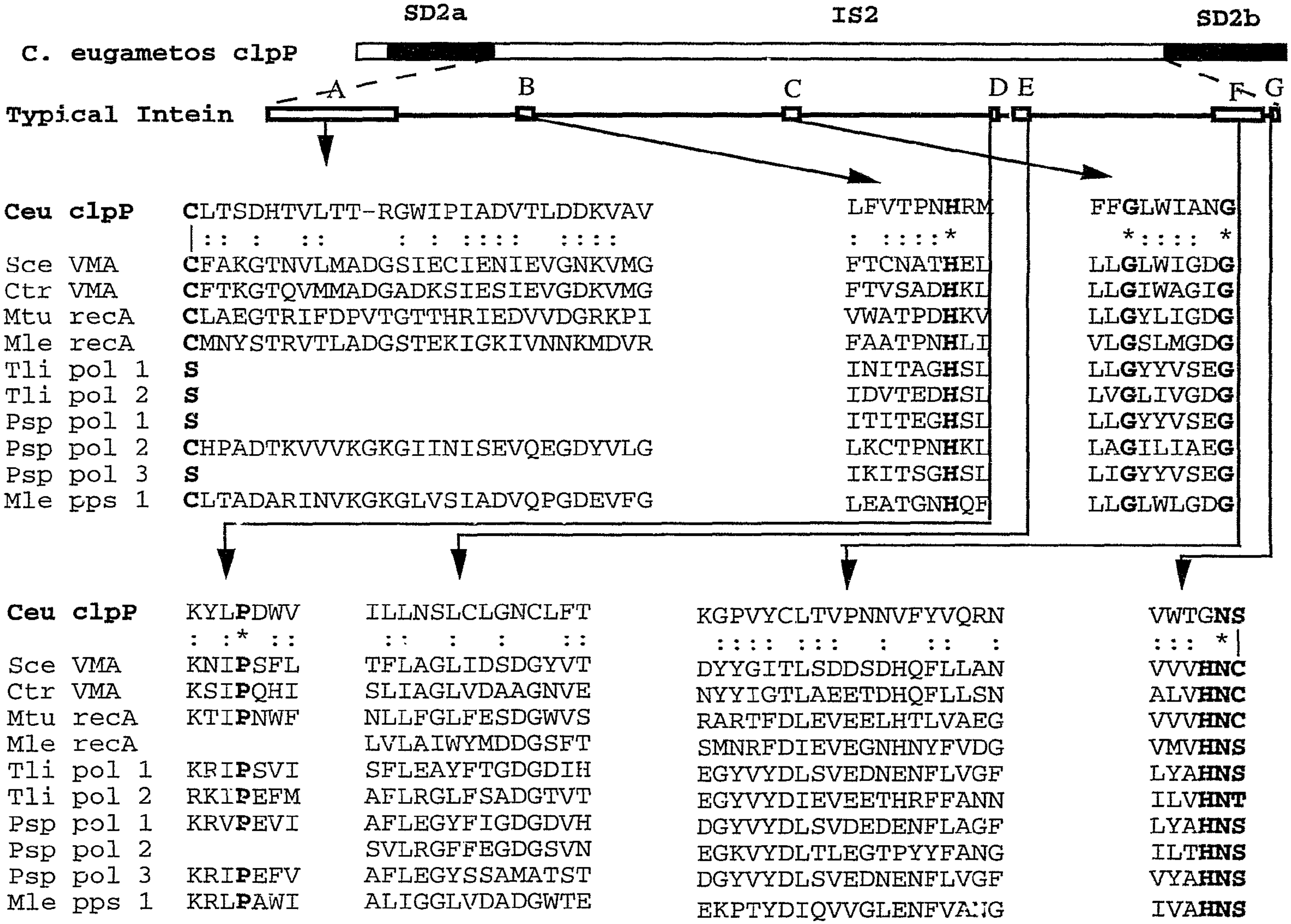
A. Identification of IS2 of *Chlamydomonas eugametos* ClpP as a degenerate intein that can be restored to protein splicing by a single amino acid substitution

a). Identification of intein-like features in the IS2 sequence

Among the several known inteins, there is little or no overall sequence similarity, except between inteins in the same protein of related organisms. However, several short sequence blocks have been recognized that show a low, but statistically significant, level of conservation among the known inteins (Pietrokovski, 1994). There are seven such sequence blocks in total, including sequence blocks A located at the N terminus of the intein, sequence block G at the C terminus of the intein, and five sequence blocks (B to F) that are located internally. Sequence block D is not present in all intein sequences, and sequence block A is present only when the intein starts with a Cys residue.

A detailed analysis of the IS2 sequence revealed that it contains sequence blocks similar to each of the seven sequence blocks conserved in known inteins (Fig. 5-3), although there is no overall sequence similarity between IS2 and any of the known inteins. Furthermore, these sequence blocks in IS2 are positioned similarly to their corresponding sequence blocks in known inteins, and they also contain several residues that are universally conserved in the corresponding sequence blocks of all known inteins. As shown in Fig. 5-3, the N terminus of *Chlamydomonas eugametos* (Ceu) IS2 is C (cysteine). The N terminus of the C-terminal flanking sequence is S (serine), which is similar to Mle RecA, Pps Pol 2, and Mle pps 1. The H (histidine) in block B, Gs (glycine) in block C, and P (proline) in block D are all present in the Ceu IS2 in the similar blocks, although the functional relationship of these amino acid residues to the protein splicing activity is currently unknown. Notably, the amino acid residue at position 4 in block G is

Figure 5-3. The alignment of *Chlamydomonas eugametos* ClpP IS2 sequence with those of inteins or intein-like sequences. The hatched box represents the IS2 of *Chlamydomonas eugametos* ClpP, open boxes are conserved sequence blocks found in a typical intein, which are also present in the Ceu IS2 at similar positions. So far, 10 intein-like sequences have been identified from three different organisms. In yeast, an intein was found in the vascular membrane subunit of the ATPase gene in *Saccharomyces cerevisiae* (Sce VMA) and *Candida tropicalis* (Ctr VMA). In mycobacteria, it was identified in the *recA* gene of *Mycobacterium tuberculosis* (Mtu *recA*) and *Mycobacterium leprae* (Mle *recA*), and the pps 1 open reading frame of *Mycobacterium leprae* (Mle pps 1). The DNA polymerase of the extreme thermophilic archaebacterium *Thermococcus litoralis* contains two intein-like sequences (Tli pol 1 and Tli pol 2), and that of *Pyrococcus* species strains GB-D and KOD1 contains three (Psp pol 1, Psp pol 2, and Psp pol 3). The highlighted amino acid residues in block A represent the first amino acid residue of the inteins, the last one in box G represents the first amino acid residue of the C-terminal flanking sequence domains. Amino acid residues that are identical in all the inteins known so far are also highlighted. Only the amino acid sequences from those conserved sequence boxes are aligned; the rest of the sequence, which shows no similarity to any of the inteins listed here, is omitted. Double dots represent the amino acid residues in Ceu IS2 that are identical or similar to the corresponding amino acid residues in some of the intein sequences listed below, stars represent the amino acid residues that are identical to the corresponding amino acid residues in all the intein sequences, and these conserved amino acid residues are highlighted in bold case.



Ceu clpP	**C**LTSDHTVLTT-RGWIPIADVTLDDKVAV	LFVTPN**H**RM	FF**G**LWIAN**G**
	|:: : :: : : :::: ::::	: ::::*	*:::: *
Sce VMA	**C**FAKGTNVLMADGSIECIENIEVGNKVMG	FTCNAT**H**EL	LL**G**LWIGD**G**
Ctr VMA	**C**FTKGTQVMMADGADKSIESIEVGDKVMG	FTVSAD**H**KL	LL**G**IWAGI**G**
Mtu recA	**C**LAEGTRIFDPVTGTTHRIEDVVDGRKPI	VWATPD**H**KV	LL**G**YLIGD**G**
Mle recA	**C**MNYSTRVTLADGSTEKIGKIVNNKMDVR	FAATPN**H**LI	VL**G**SLMGD**G**
Tli pol 1	**S**	INITAG**H**SL	LL**G**YYVSE**G**
Tli pol 2	**S**	IDVTED**H**SL	LV**G**LIVGD**G**
Psp pol 1	**S**	ITITEG**H**SL	LL**G**YYVSE**G**
Psp pol 2	**C**HPADTKVVVKGKGIINISEVQEGDYVLG	LKCTPN**H**KL	LA**G**ILIAE**G**
Psp pol 3	**S**	IKITSG**H**SL	LI**G**YYVSE**G**
Mle pps 1	**C**LTADARINVKGKGLVSIADVQPGDEVFG	LEATGN**H**QF	LL**G**LWLGD**G**

Ceu clpP	KYL**P**DWV	ILLNSLCLGNCLFT	KGPVYCLTVPNNVFYVQRN	VWTG**NS**
	: :* ::	:: : : ::	:::: ::: : :: :	::: *|
Sce VMA	KNI**P**SFL	TFLAGLIDSDGYVT	DYYGITLSDDSDHQFLLAN	VVV**HNC**
Ctr VMA	KSI**P**QHI	SLIAGLVDAAGNVE	NYYIGTLAEETDHQFLLSN	ALV**HNC**
Mtu recA	KTI**P**NWF	NLLFGLFESDGWVS	RARTFDLEVEELHTLVAEG	VVV**HNC**
Mle recA		LVLAIWYMDDGSFT	SMNRFDIEVEGNHNYFVDG	VMV**HNS**
Tli pol 1	KRI**P**SVI	SFLEAYFTGDGDIH	EGYVYDLSVEDNENFLVGF	LYA**HNS**
Tli pol 2	RKI**P**EFM	AFLRGLFSADGTVT	EGYVYDIEVEETHRFFANN	ILV**HNT**
Psp pol 1	KRV**P**EVI	AFLEGYFIGDGDVH	DGYVYDLSVDEDENFLAGF	LYA**HNS**
Psp pol 2		SVLRGFFEGDGSVN	EGKVYDLTLEGTPYYFANG	ILT**HNS**
Psp pol 3	KRI**P**EFV	AFLEGYSSAMATST	DGYVYDLSVEDNENFLVGF	VYA**HNS**
Mle pps 1	KRL**P**AWI	ALIGGLVDADGWTE	EKPTYDIQVVGLENFVANG	IVA**HNS**

Figure 5-3

H (histidine) in all the inteins, but it is G (glycine) in Ceu IS2. In yeast Sce VMA, the protein splicing activity was completely abolished when this H was mutagenized to G, suggesting that this H is relevant to the protein splicing activity (Cooper et al, 1993) and probably directly participates in the protein splicing process.

b) Characterization of the unmodified (wildtype) IS2 sequence

Identification of the seven intein-related sequence blocks in the IS2 sequence raised the possibility that IS2 may be also an intein that could be removed from the ClpP precursor protein through protein splicing. This possibility was first examined *in vivo* by analyzing the protein product of the *clpP* gene in *Chlamydomonas eugametos*. Total cellular proteins were extracted from *Chlamydomonas eugametos* cells grown logarithmically in liquid culture, resolved by SDS-polyacrylamide gel electrophoresis, and blotted onto nitrocellulose membrane. The membrane was then subjected to standard Western blots using antibodies raised against ClpP protein sequence domains of *Chlamydomonas reinhardtii* (described below). However, none of these antibodies detected any protein band among the total cellular protein of *Chlamydomonas eugametos* (data not shown), indicating that a ClpP protein is not accumulated to a detectable level under the conditions studied.

The possibility of IS2-associated protein splicing was then investigated in *E. coli* cells. The *Chlamydomonas eugametos clpP* gene, including the IS1-coding sequence, was amplified by PCR from *Chlamydomonas eugametos* DNA and cloned into the expression vector pET, placing it behind a T7 promoter. The resulting recombinant plasmid was transformed into *E. coli* strain DE3 cells, which have an IPTG-inducible T7 RNA polymerase gene. These *E. coli* cells were grown in liquid culture and induced to express the *clpP* gene by the addition of IPTG. After the induction, total cellular proteins were resolved on an analytical SDS-polyacrylamide gel and stained with Coomassie Blue.

Only a protein band corresponding to the unprocessed ClpP precursor protein was induced (data not shown), indicating that the IS2 sequence is not removed from the precursor protein. It was therefore concluded that the unmodified (wildtype) IS2 sequence does not support protein splicing in *E. coli* cells.

c) Site-directed mutagenesis of the IS2 sequence

The fact that IS2 sequence contains intein-like features suggested that it is an intein, but its inability to support protein splicing in *E. coli* cells suggested that it may have accumulated critical mutations that abolished its protein splicing activity. One such critical mutation may be the Gly residue located in the sequence block G near the C-terminus of IS2, since in all known inteins this position is occupied by a His residue. In the yeast VMA1 intein, replacement of this His residue by a Gly residue abolished the intein's protein splicing activity. It was therefore reasoned that the IS2 might be restored to protein splicing by changing the Gly residue to a His residue through site-directed mutagenesis.

The mutagenesis was carried out by using a PCR-mediated site-directed method, in which synthetic oligonucleotides containing the desired mutations are used as PCR primers to amplify the target DNA. The amplified DNAs, now containing the desired mutations, are cloned into appropriate plasmid vectors. DNA sequence determination is then used to confirm the results of the mutagenesis. By using this PCR-mediated site-directed mutagenesis method, the glycine codon (GGT) was changed into a histidine codon (CAT) (Fig. 5-4). DNA sequence determination of the mutagenized DNA confirmed that the desired change was made, and no unwanted mutation was introduced in this process.

Figure 5-4. Mutagenesis of *Chlamydomonas eugametos* IS2. Hatched box represents the Ceu IS2 sequence, the DNA sequence at the C-terminal of IS2 and the corresponding amino acid sequence are shown below. Mutation introduced by PCR-mediated site-directed mutagenesis was at the 3-terminal junction, where the underlined GGT (encoding glycine) was changed into CAT (encoding histidine). Located 140 bp upstream of that site is an Spe I site (ACTAGT), when it is cut by Spe I and the resulting ends filled in prior to religation, four bases are inserted (CTAG). The underlined TAG in the insertion site is in-frame with the upstream coding sequence, thus serving as a termination codon (AMB).

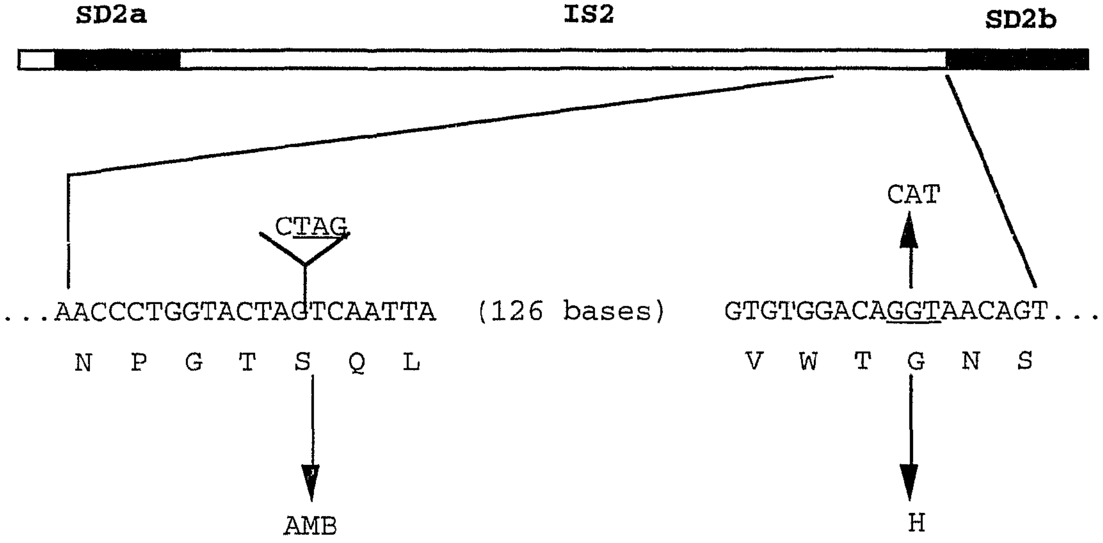

SD2a
IS2
SD2b
CTAG
CAT
...AACCCTGGTACTAGTCAATTA (126 bases) GTGTGGACAGGTAACAGT...
N P G T S Q L
V W T G N S
AMB
H

Figure 5-4

Another mutation was also generated in the IS2 coding sequence by creating a translation termination codon within this sequence, in order to distinguish protein splicing from RNA splicing. A fundamental difference between protein splicing and RNA splicing is that intein must be translated. Creating the termination codon should definitely prevent protein splicing but should not normally affect RNA splicing. In order to create an in-frame termination codon in the IS2-coding sequence, this DNA was cut at a unique Spe I site located 147-bp upstream of the 3'-end of IS2 coding sequence. This cleavage leaves a pair of cohesive ends with a 5'-end 4-bp overhang (CTAG), which was filled in and the resulting blunt ends ligated. This process creates a 4-bp insertion at the Spe I site and introduces a termination codon (TAG) at this same location (Fig. 5-4).

d) Characterization of the modified IS2 (IS2-m) for its protein splicing activity

The site-directed mutagenesis described above created a modified IS2 sequence (IS2-m) in which a Gly residue is replaced by a His residue. In order to find out whether this modified IS2 can support protein splicing in *E. coli* cells, four recombinant plasmids were constructed (Fig. 5-5). For each recombinant plasmid, a 2-kbp BamH I-Xho I DNA fragment from the *Chlamydomonas eugametos clpP* gene was inserted into the expression vector pMAL between its BamH I and Sal I sites, creating a continuous open reading frame for a fusion protein consisting of a portion of the ClpP protein sequence fused to a maltose-binding protein sequence. The partial ClpP sequence is the C-terminal two-thirds of the complete ClpP sequence of *Chlamydomonas eugametos*, which consists of the complete IS2 sequence plus approximately 100 residues of its N-terminal flanking sequence and approximately 100 residues of its C-terminal flanking sequence. Plasmid pWC173-1 contains the unmodified (wild-type) IS2 sequence, while plasmid pWC173-2 contains the modified IS2 sequence (IS2-m) with the Gly to His change. In plasmid

Figure 5-5. Clones of wild type, mutant, and disrupted Ceu IS2 in pMAL vector. The four clones used in the expression of Ceu IS2 and related proteins are summarized. Shown on the top is the Ceu clpP gene and its restriction map. A BamH I-Xho I fragment was cloned in pMAL vector between BamH I and Sal I sites to get the wild-type pWC173-1. The insertion contains the whole Ceu IS2 and about 300 bp from its upstream and 300 bp from its downstream sequence domains. Glycine is encoded by the wild-type gene at the C-terminal junction of the IS2 as indicated (pWC173-1), whereas the mutant encodes a histidine at this point (pWC173-2). Disruption of the wild-type gene yields clone pWC174-1, whereas disruption of the mutant yields clone pWC174-2, the in-frame termination codon being contained within the four inserted nucleotides and underlined. All four clones are in pMAL vector; thus, the proteins expressed from those clones would have the 42-kDa maltose-binding protein at their N-termini (the maltose-binding protein part is not shown). The predicted size of the protein expressed from each clone is given at the end of that clone, which includes the 42-kDa of a maltose-binding protein. In the mutant clone pWC173-2, if the IS2 is removed from the precursor, the resulting protein will be 66-kDa in size, of which 42-kDa would be from the maltose-binding protein, whereas 24-kDa would be from 110 amino acid residues upstream of the IS2 and 107 amino acid residues downstream of the IS2. The size of IS2 alone would be 51-kDa.

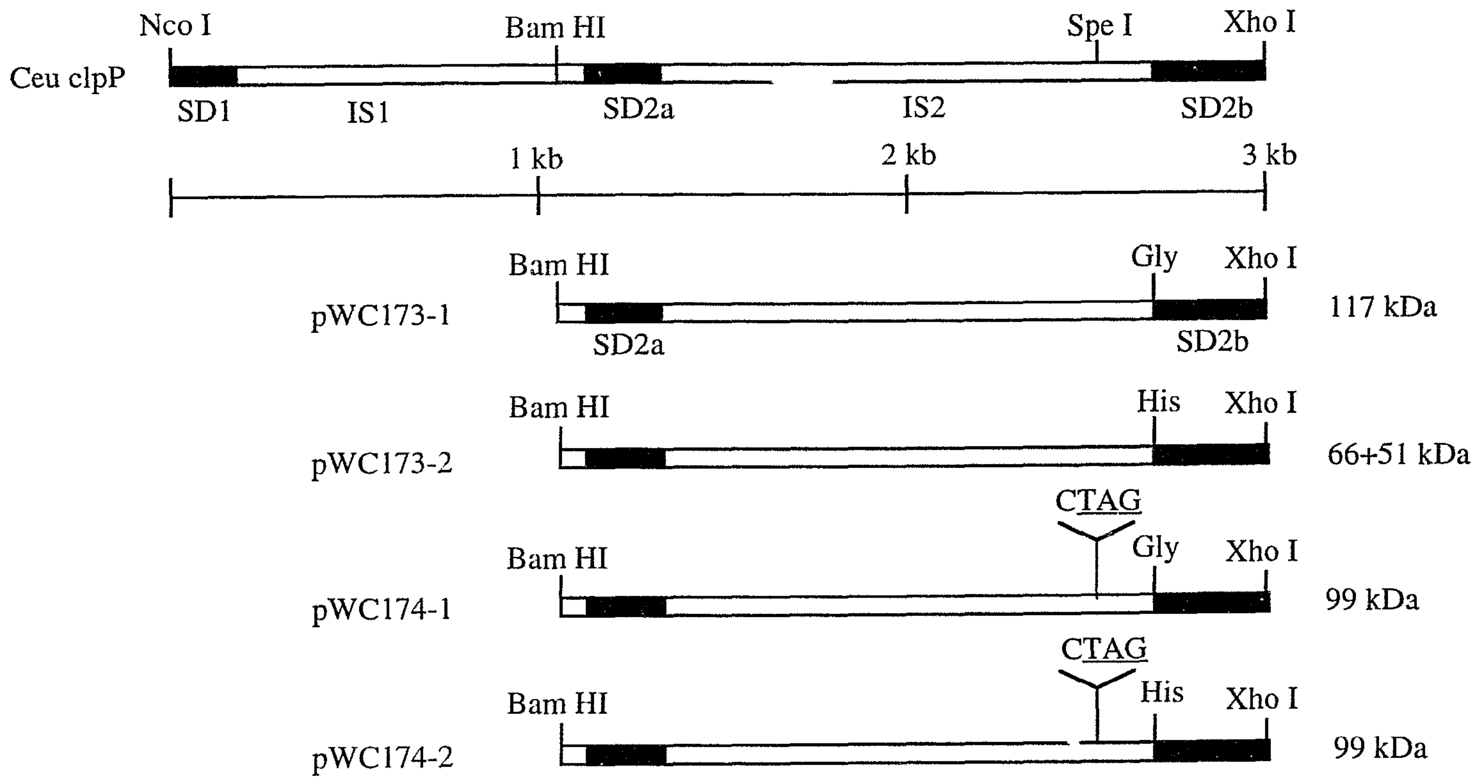

Nco I
Bam HI
Spe I
Xho I
Ceu clpP
SD1
IS1
SD2a
IS2
SD2b
1 kb
2 kb
3 kb
Bam HI
Gly
Xho I
pWC173-1
117 kDa
SD2a
SD2b
Bam HI
His
Xho I
pWC173-2
66+51 kDa
CTAG
Gly
Xho I
Bam HI
pWC174-1
99 kDa
CTAG
His
Xho I
Bam HI
pWC174-2
99 kDa

Figure 5-5

pWC174-1, the IS2 sequence contains the termination codon, while in plasmid pWC174-2, the IS2 sequence contains both the Gly to His change and the termination codon.

E. coli cells transformed with each of the above recombinant plasmids were grown in liquid culture and induced by IPTG to produce the IS2-containing fusion proteins. After the induction, total cellular proteins were resolved by analytical SDS-polyacrylamide gel electrophoresis and stained with Coomassie blue. The results are shown in Fig. 5-6. For plasmid pWC173-1 with the unmodified (wildtype) IS2 sequence, only a 117-kDa protein band was induced (lane 2), which corresponds in size to the predicted precursor protein. This result is consistent with the earlier conclusion that the unmodified (wildtype) IS? does not support protein splicing. For plasmid pWC173-2 with the modified (mutant) IS2 containing the Gly to His change, a 66-kDa protein band was produced in addition to the 117-kDa precursor protein (lane 3). The 66-kDa protein corresponds in size to a predicted spliced protein, suggesting that it was produced by protein splicing in which the IS2 sequence was removed from the precursor protein. For plasmids pWC174-1 and pWC174-2, both having a termination codon in the IS2 sequence, only a 99-kDa protein band was induced (lane 4 and 5), which corresponds in size to a truncated protein as predicted (Fig. 5-6). This result indicates that translation through the IS2 coding sequence is necessary for the production of the 66-kDa protein, further suggesting that the 66-kDa protein is a product of protein splicing rather than RNA splicing or premature translation termination.

In order to completely rule out the possibility of RNA splicing, mRNA transcript from the pWC173-2 plasmid was analyzed by the RT-PCR method. Total RNA was isolated from *E. coli* cells harboring the pWC173-2 plasmid and induced by IPTG. The RNA sample was treated with RNase-free DNase to remove any residual contaminating DNAs. Reverse transcription (RT) was then carried out by using a synthetic

Figure 5-6. The expression of IS2 in *E. coli.* A single colony was grown in 2 ml LB medium containing 50 μg/ml ampicillin at 37 °C overnight, the culture was diluted 100 times with fresh LB containing ampicillin and grow at 37 °C for 2 to 3 hrs to a cell density of 0.5 at A_{600}. IPTG was added to a final concentration of 1 mM and the cells were grown at 37 °C for 2 hrs. Cells from 1 ml culture were harvested and resuspended with 150 μl of protein loading buffer and heated in a boiling water bath for 5 min. A sample of 5 μl was loaded onto an analytical SDS gel and run at 200 V for 45 min. The gel was stained in 0.1% Coomassie blue for 15 min, then destained for 2 hrs. The MalE lane represents the total proteins induced from the pMAL vector with a disruption at the cloning linker as a control, where only the 42-kDa maltose-binding protein is induced. Lanes Wild-type, Mutant, Wild-type Truncated and Mutant Truncated are proteins induced from clone pWC173-1, pWC173-2, pWC174-1 and pWC174-2, respectively. Protein size markers (Promega, middle range) are shown at the left side, sizes of induced proteins are shown at the right side.

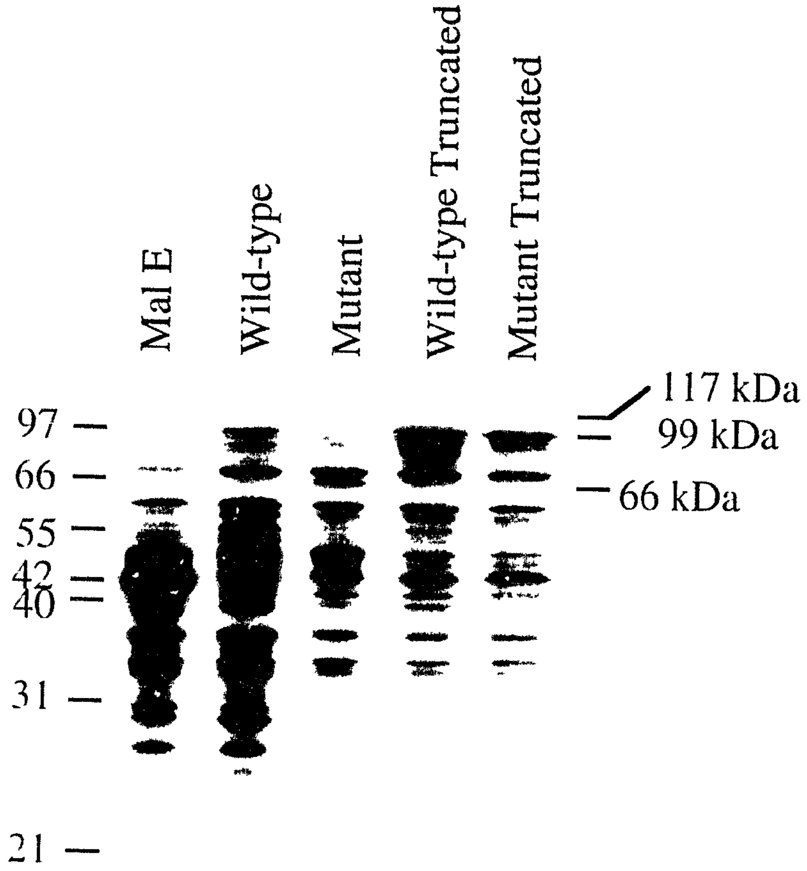

Mal E
Wild-type
Mutant
Wild-type Truncated
Mutant Truncated
97
66
55
42
40
31
21
117 kDa
99 kDa
66 kDa

Figure 5-6

oligonucleotide primer designed specifically for mRNA of the plasmid-encoded and IS2-containing fusion protein. The resulting cDNA was amplified by PCR, using a pair of oligonucleotide primers specific for the target cDNA. As shown in Figure 5-7, only the unspliced mRNA was detected (right lane) by this method. Nothing was seen in the control lane (middle lane), which contains everything as in the right lane except the reverse transcriptase during cDNA synthesis. The result indicates that the band seen in the right lane was amplified from cDNA, not from plasmid DNA contamination. In conclusion, there is no spliced mRNA detectable by the RT-PCR method.

e). Purification of the 66-kDa protein as a putative spliced protein

In order to further investigate the identity of the 66-kDa protein produced from the recombinant plasmid pWC173-2, this protein needs to be purified from the corresponding *E. coli* cells. If the 66-kDa protein is indeed the spliced protein, it should contain the maltose-binding protein at its N-terminus and bind to an amylose affinity column. This was shown to be the case (Fig. 5-8). When the *E. coli* cell lysate was passed through an amylose-Sepharose column, the 66-kDa protein was specifically retained by the column, along with the 117-kDa precursor protein as predicted (purified lane). Several other minor proteins were also retained by the column (purified lane), but it is not clear whether they represent protein splicing intermediates or protein degradation products. This affinity chromatography method, however, was not suitable for purifying large amount of the 66-kDa protein for three reasons. First, most of the 66-kDa protein produced in the *E. coli* cells was insoluble (insoluble lane), indicating that it was trapped in inclusion bodies. Second, although the protein can be isolated by amylose affinity column chromatography, several nonspecific proteins were co-purified. And third, the 117-kDa precursor protein can not be separated from the 66-kDa protein (purified lane).

Figure 5-7. Reverse transcription of *clpP* transcript and PCR amplification of cDNA. The clone pWC173-2 was grown until A_{600} was 0.5 and transcription of the gene was induced by the addition of IPTG at this point and the cell were allowed to grow for 30 min at 37 °C. Total RNA was extracted and digested with RNase-free DNase as described in Materials and Methods. Total RNA (5 µg) was reverse transcribed and cDNA was amplified by PCR. Sample in the midle lane is the reaction that the reverse transciptase was omitted so that anything amplified by the subsequent PCR will be from plasmid contamination. Nothing , as amplified in the reaction corresponding to this lane, suggesting that DNA contamination was negligible. Only the 2.1 kbp unspliced mRNA was detected (right lane), whereas the predicted size of a spliced mRNA should be 0.7 kbp. The left lane is lambda DNA cut with Hind III, serving as DNA molecular marker. Size of each marker band (from top to bottom) is 23.1, 9.4, 6.6, 4.4, 2.3, 2.0 and 0.6 kbp, respectively.

Figure 5-7

Figure 5-8. Purification of ClpP-related proteins by amylose affinity column chromatography. A single colony was grown in 2 ml LB medium containing 50 μg/ml of ampicillin overnight at 37 °C with vigorous shaking. The 2 ml overnight culture was diluted into a 100 ml fresh LB medium containing 50 μg/ml of ampicillin and the culture was grown for 2 to 3 hrs until the absorbance at A_{600} was 0.5. One fraction was taken at this point as the uninduced control. IPTG was then added and the cells were grown for another 2 hrs. after which another fraction was taken as the induced total cellular protein. Cells from the remaining culture were harvested by centrifugation and resuspended in column buffer. After sonication and centrifugation, the pellet was resuspended in 1x protein loading buffer (insoluble fraction), and a portion of the supernatant was taken as the soluble fraction. The supernatant was then subjected to column chromatography as described in the Materials and Methods section. A fraction was taken after purification as the flow-through fraction, and the protein eluted from the column was used as the pure protein fraction.

The uninduced and induced lanes are total proteins before and after the addition of IPTG, the insoluble lane is sample from the pellet, soluble lane is from the supernatant, the flow-through and the column-purified lanes are as indicated. Protein loaded onto each lane was adjusted so that each lane contains proteins from the same number of cells in respect to the start culture (except in the purified lane). Left lane is the protein marker (middle range, Promega), with the sizes of each protein standard indicated.

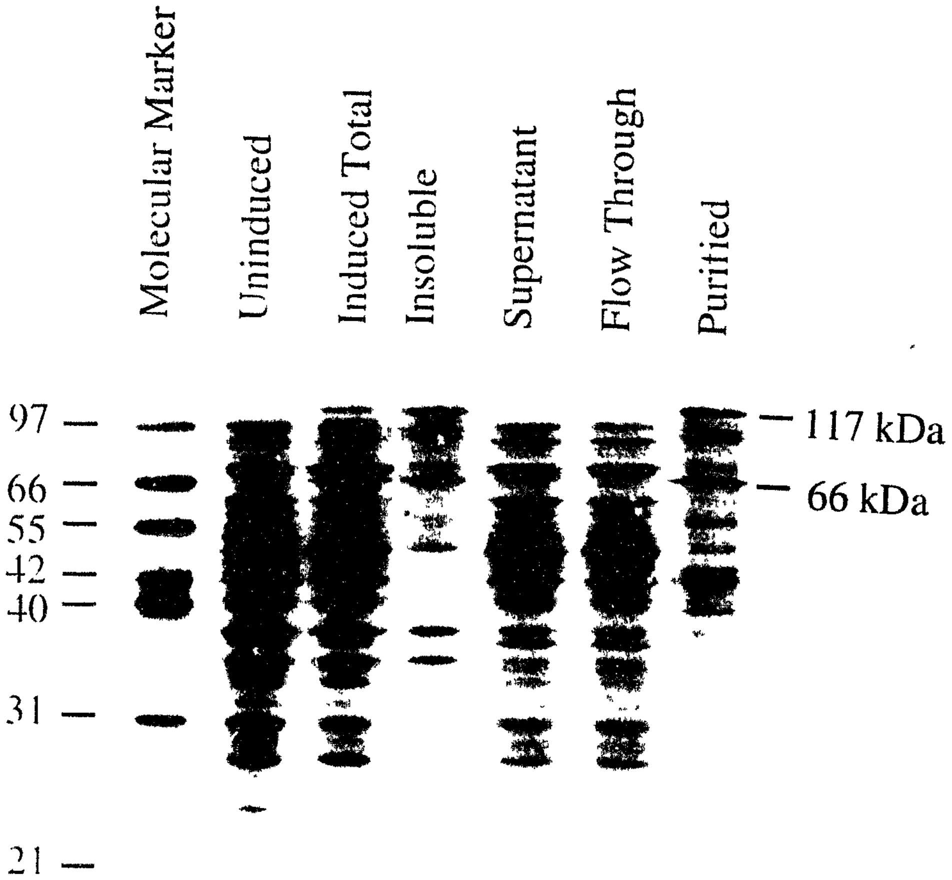

Molecular Marker
Uninduced
Induced Total
Insoluble
Supernatant
Flow Through
Purified
97
66
55
42
40
31
21
117 kDa
66 kDa

Figure 5-8

Fortunately, the 66-kDa protein migrates on an SDS-polyacrylamide gel to a region containing no other abundant protein, as judged by the protein band pattern in the total protein lane (compare the insoluble lane to other lanes). Excision of a gel slice containing the 66-kDa protein band followed by electroelusion could be a more effective way of purifying the protein. Based on these observations, the 66-kDa protein was purified from the insoluble fraction of the *E. coli* cell lysate (insoluble lane) by excising the 66-kDa protein band from a stained gel and electroeluting the protein from it.

f). Identification of the 66-kDa protein as a spliced protein by micro-protein sequencing

The 66-kDa protein purified by electroelution from a gel slice was subjected to protein structure analysis in order to confirm its identity as a spliced protein. This protein, produced from the pMAL-based pWC173-2 recombinant plasmid, was predicted to be a fusion protein consisting of the ClpP sequence fused to a maltose-binding protein at its N terminus. A protease Factor Xa cleavage site (Ile-Glu-Gly-Arg) is located between the ClpP sequence and the maltose-binding protein sequence. As expected, treatment of the 66-kDa protein with the protease Factor Xa resulted in two smaller polypeptides. One is 42-kDa in size, corresponding to the maltose-binding protein. The other one is 24-kDa in size, which corresponds to the predicted ClpP sequence (Fig. 5-9). It was also noted that cleavage of the 66-kDa protein did not proceed to completion. After 2 hrs of treatment with the protease Factor Xa, approximately one-third of the 66-kDa protein still remained (see the 2 hours lane). This could be a result of incomplete cleavage of the 66-kDa fusion protein, but the remaining 66-kDa protein could represent an unrelated protein that co-purifies with the fusion protein. Longer periods of treatment did result in complete disappearance of the 66-kDa protein, but this also resulted in the appearance of apparently nonspecific cleavage products, without increasing the proportion of the 24-kDa polypeptide.

Figure 5-9. The digestion of 66-kDa protein with protease Factor Xa. Lane 1 is the 66-kDa protein before digestion, lanes 0 to 2 hours are the digestion mixture (5 µg of 66-kDa protein, 0.1 µg of factor Xa, in 1x Factor Xa buffer, see Materials and Methods section for details) taken at 0 min, 30 min, 1 hr and 2 hrs of incubation, the right lane was taken from bulk cutting mixture (10 µg of 66-kDa protein, 1 µg of Factor Xa, in 1x Factor Xa buffer) after 2 hrs incubation. Each lane contains 1 µg of 66-kDa protein as the starting material. Protein size markers (Promega, middle range) are shown at the left side, the 66-kDa substrate, the 42-kDa maltose-binding protein, the 24-kDa protein, which was sequenced subsequently, and the Factor Xa protein band, are indicated at the right side.

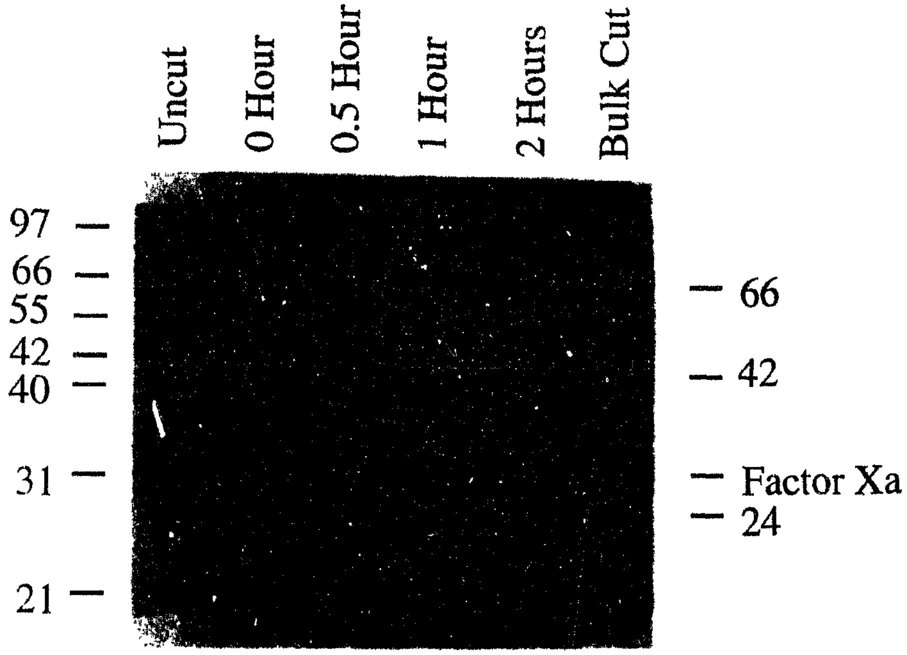

Uncut
0 Hour
0.5 Hour
1 Hour
2 Hours
Bulk Cut
97
66
55
42
40
31
21
66
42
Factor Xa
24

Figure 5-9

A sufficient amount of the 24-kDa polypeptide was then prepared in order to carry out micro-protein sequencing. Approximately 10 µg of the 66-kDa protein was treated with 1 µg of the protease Factor Xa for 2 hrs. The resulting polypeptide fragments were resolved on an SDS-polyacrylamide gel, blotted onto a PVDF membrane, and stained briefly by Ponceau S to visualize the polypeptide bands. The 24-kDa polypeptide band was excised ar·l sent to the Harvard University Microchemistry Facility for further analysis as described below.

First, the 24-kDa polypeptide was treated with protease trypsin to cleave it into smaller pieces suitable for protein sequencing. This was also necessary in order to isolate a smaller fragment whose sequence spans the predicted splice junction of the protein. After the trypsin treatment, the resulting peptide fragments were resolved by FPLC (Fig. 5-10). Four fragments (#5, #8, #9, #17) were selected as prime targets based on their absorption spectrum. These fragments were subjected to mass spectrometric analysis in order to determine their accurate masses. The results from this analysis strongly indicated that each fragment corresponded to a piece of the predicted sequence of a spliced protein, and more importantly, fragment number 9 corresponded to a piece of sequence spanning the predicted spliced junction.

Fragment number 9 was therefore chosen for micro-protein sequencing. This revealed a 30-residue sequence that matched perfectly with the predicted sequence spanning the splice junction. This clearly demonstrated that the IS2-m sequence was precisely ιemoved from the precursor protein, and the flanking sequences were joined together by a bona fide peptide bond. These results, together with the various findings described above, clearly identified the 66-kDa protein as a spliced protein produced

Figure 5-10. FPLC profile. The deduced protein sequence from the DNA sequence cloned in pWC173-2, with the precise removal of the IS2 sequence, is shown on the top. After digestion with trypsin (cuts after Lys and Arg residues), 20 polypeptides are expected to be obtained (underlined and numbered from 1 to 20). These polypeptides were resolved on FPLC chromatography (profile at the bottom). Four polypeptides were chosen based on the content of their aromatic amino acid residues, and the molecular masses were determined. One of them has the closest molecular mass to the polypeptide 9 (3829.8 versus 3832.2), so this one was chosen to be sequenced.

```
ISEFGSLNYLDFYSYNDSYNDFK TAPR GK QAER AFQEEESK K VFVIINSFGGSVGNGITVHDALQFIK
           1             2   3   4      5     6              7

AGSLTLALGVAASAASLALAGGTIGER YVTEGCHVMIHQPESsIQGQASDIWIDSQEIMK IR LDVAE
             8                              9                 10   11

IYSLATYR PR HK ILR DLDK DFYLTATETIHYGLADEIASNEVMQEIIEMTSK VWDYHDTK QQR LLESR
         12 13 14   15                 16                      17     18   19

DSTTSGADTQSQN        5: 967.0       8: 2398.7       9: 3832.2       17: 1063.1
      20                963.9          2399.5          3829.8           1064.0
```

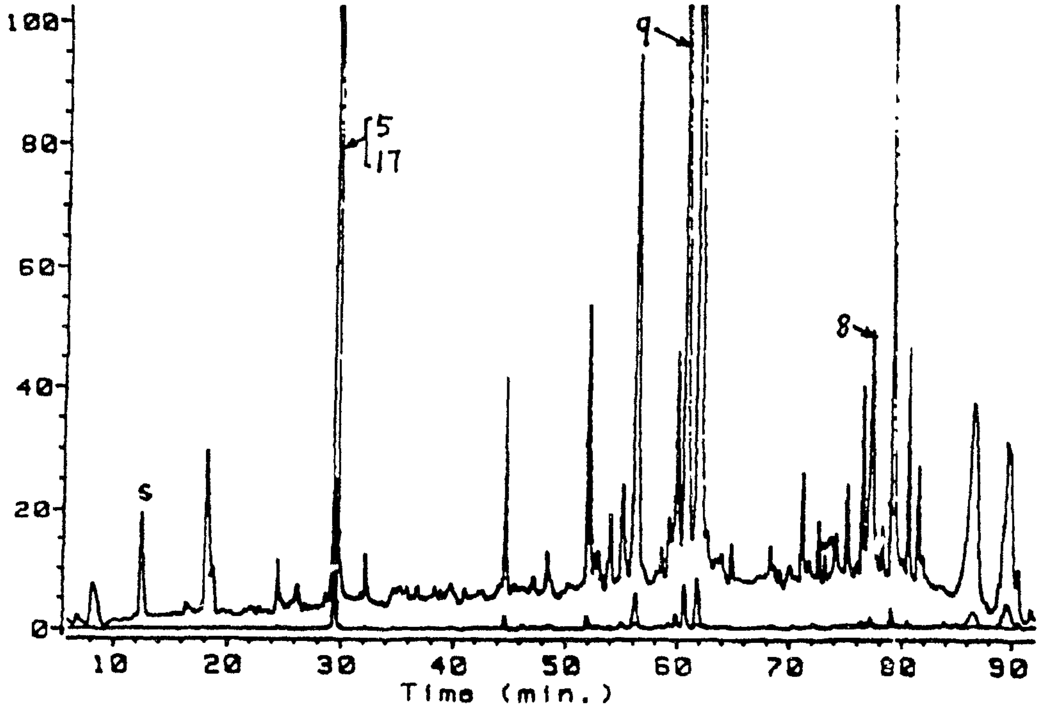


Figure 5-10

through protein splicing. The IS2-m sequence was thus identified as an active intein, and the unmodified (wildtype) IS2 is therefore a degenerate intein that can be restored to protein splicing by a single amino acid substitution.

B. Characterization of the IS1-containing ClpP protein of *Chlamydomonas reinhardtii*

The expression and protein product of *Chlamydomonas reinhardtii clpP* gene was also investigated. It was investigated first in *E. coli* cells for possible post-translational processing such as protein splicing. The complete *clpP* gene, including the IS1-coding sequence, was amplified by PCR from *Chlamydomonas reinhardtii* DNA and subsequently cloned into the expression vector pET, placing it behind a T7 promoter. The resulting recombinant plasmid was transformed into *E. coli* strain DE3 cells which has an IPTG-inducible T7 RNA polymerase gene. These *E. coli* cells were grown in liquid culture and induced to express the *clpP* gene by the addition of IPTG. After the induction, total cellular proteins were resolved on an analytical SDS-polyacrylamide gel and stained with Coomassie Blue. Only a 59.5-kDa protein band was induced (data not shown), which corresponded in size to the unprocessed precursor protein of ClpP (Fig. 5-1). It was concluded that there was no processing of the *Chlamydomonas reinhardtii* ClpP precursor protein in *E. coli.*

The expression of this *clpP* gene and possible post-translational processing were then investigated in *Chlamydomonas reinhardtii* chloroplast. First, antibodies specific to different regions of the gene product were raised for Western blotting analysis. As shown in Fig. 5-11, two DNA fragments from the *clpP* gene were cloned separately into the expression vector pMAL: an EcoR I-Nde I fragment encoding 54 amino acid residues from the N-terminus of the ClpP protein (designated as N domain), anc Hind III-EcoR I fragment encoding 104 amino acid residues from the C-terminus of the ClpP protein (designated as C domain). Each of the two DNA fragments was placed behind and in-frame with the pMAL vector-encoded maltose-binding protein, so that the corresponding N domain or C domain sequences can be produced as part of a larger fusion protein. The resulting recombinant plasmids were transformed into *E. coli* cells, and the transformants were subsequently induced by IPTG to produce fusion proteins containing the corresponding ClpP protein fragments. The fusion proteins were resolved by SDS-polyacrylamide gel electrophoresis, purified by electroelution from excised gel slices, and used as antigens to raise antibodies in rabbits. The resulting antibodies, designated respectively as N antibody and C antibody, were affinity-purified by absorption onto antigen-impregnated cellulose membranes and shown to be specific for the corresponding N-domain and the C-domain polypeptides. In addition, the whole ClpP protein produced in *E. coli* cells containing the pET-derived plasmid (see above) was also used as antigen to raise a rabbit antibody, which was designated P antibody. Specificity test of the P antibody indicated that most of the P antibody activity was directed against the IS1 sequence (data not shown).

For Western blotting analysis, total cellular proteins were extracted from *Chlamydomonas reinhardtii* cells grown logarithmically in liquid culture, resolved by SDS-polyacrylamide gel electrophoresis, and blotted onto nitrocellulose membrane. The membrane was subjected to standard Western blot using N, C, and P antibodies separately. Each of the three antibodies detected only the 59.5-kDa precursor protein (data not shown). It was therefore concluded that the IS1 sequence was not removed from the precursor protein in *Chlamydomonas reinhardtii* chloroplasts, at least under the growth conditions used in this study.

Figure 5-11. Subcloning of *Chlamydomonas reinhardtii clpP* gene. A schematic comparison of the *Chlamydomonas reinhardtii* ClpP and conventional ClpP proteins is shown on top. The black boxes are protein sequence regions similar to each other, the open box is the insertion sequence (IS1), which is not similar to any sequences known. A restriction map of the *Chlamydomonas reinhardtii* clpP gene is shown in the middle. The EcoR I-Nde I fragment corresponds to the N-terminus of the *C. reinhardtii* ClpP protein, which encodes a polypeptide 54 amino acid residues long; the Hind III-EcoR I fragment corresponds to the C-terminus of the *C. reinhardtii* ClpP protein, which encodes a polypeptide 104 amino acid residues long. These two fragments were cloned into pMAL vector. The whole gene was cloned into pET vector. Proteins expressed from pMAL vector are fusion proteins with a 42-kDa maltose-binding protein at the N terminus, whereas protein expressed from the pET vector is a non-fusion protein. Those proteins were purified from *E. coli* and injected into rabbits to raise antibodies. Resulting antibodies were designated N-, C-, and P antibodies respectively. The predicted size of the whole *C. reinhardtii* ClpP is 59.5-kDa, as indicated.

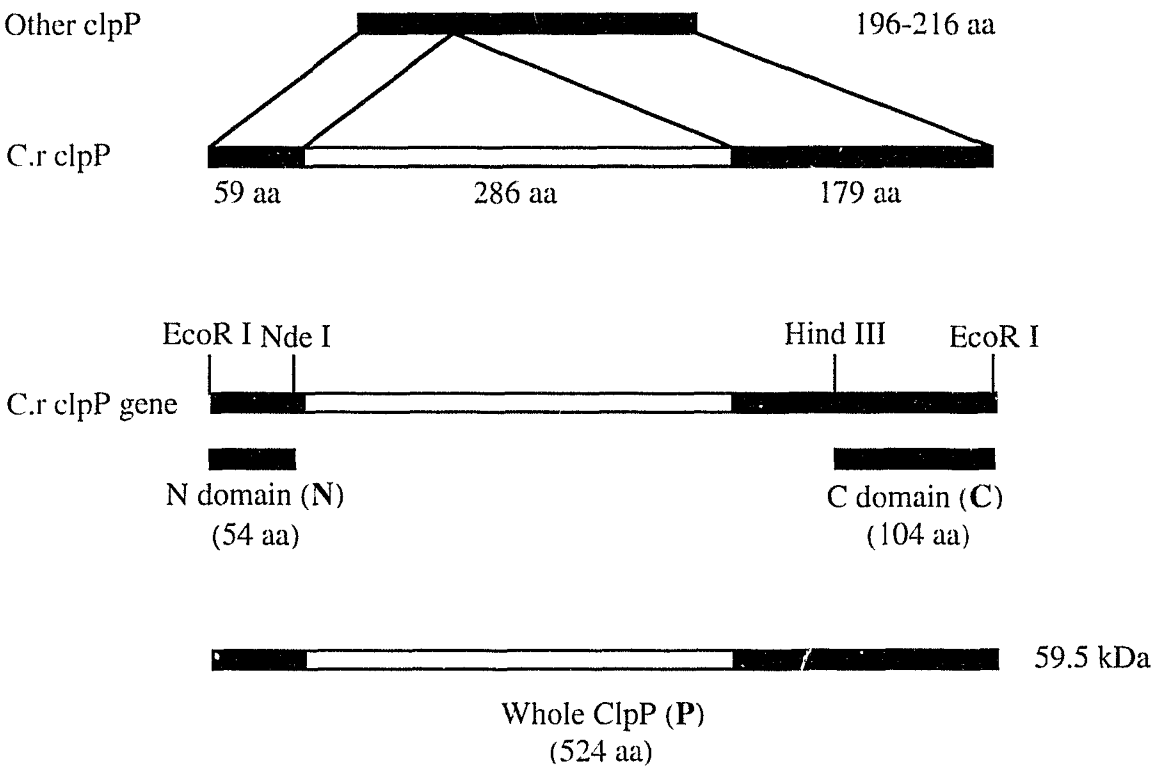

Other clpP
196-216 aa
C.r clpP
59 aa
286 aa
179 aa
EcoR I
Nde I
Hind III
EcoR I
C.r clpP gene
N domain (**N**)
(54 aa)
C domain (**C**)
(104 aa)
59.5 kDa
Whole ClpP (**P**)
(524 aa)

Figure 5-11

C. Characterization of the *rps3* gene product of *Chlamydomonas reinhardtii*

The chloroplast *rps3* gene of *Chlamydomonas*, like the *clpP* gene, appear to contain a large translated insertion sequence (Fig. 5-2). However, unlike the IS2 sequence of ClpP, the insertion sequence of Rps3 does not have any recognizable sequence motifs that are similar to the conserved sequence blocks of known inteins. Furthermore, the *Chlamydomonas reinhardtii* Rps3 protein, when produced in *E. coli* cells, did not undergo protein splicing or any other visible forms of processing (data not shown). In order to characterize the *rps3* gene product (or products) inside *Chlamydomonas reinhardtii* cells, antibodies were raised against specific sequence domains of the Rps3 protein, and Western blot analysis was subsequently carried out on *Chlamydomonas reinhardtii* total proteins and on isolated 30S ribosome subunit. These experiments and their results are described below.

a) Production of domain-specific antibodies

In order to prepare antigens for raising domain-specific antibodies, corresponding fragments of the Rps3 protein were produced in recombinant *E. coli* cells as parts of larger fusion proteins. This was achieved by cloning corresponding fragments of the *rps3* gene into the expression vector pMAL. As shown in Fig. 5-12, the Nco I-Nde I fragment encodes 86 amino acid residues from N-terminus of the Rps3 protein. It contains the conserved regions I and II of a conventional Rps3 protein, and was designated N domain. A Dde I-Bgl II fragment encodes amino acid residues from 346 to 488 of the Rps3 protein (143 aa long), containing conserved sequence block b and part of the block c identified in all the *Chlamydomonas* Rps3 proteins. Since it is from the middle spacer region of the *rps3* gene, it was designated S domain (spacer domain). The C domain was assigned to a Hind III-Dde I fragment, which encodes the C-terminal 01 amino acid

Figure 5-12. Subclones of *Chlamydomonas reinhardtii rps3* gene fragments. The *rps3* gene was PCR amplified from *Chlamydomonas reinhardtii* total DNA, and digested with respective restriction endonucleases. The Nco I-Nde I, Dde I-Bgl II and Hind III-Dde I fragments were isolated from a TAE buffered 0.8% agarose gel and purified by the Gene Clean^{TM} method. Each fragment was then cloned into pMAL vector at a cloning site where the insert is in frame with the upstream maltose-binding protein gene. Proteins expressed from those subclones were purified and used to raise domain-specific antibodies. Black boxes represent conserved regions of a conventional Rps3 protein, which are also present in *E. coli* and *Chlamydomonas reinhardtii* Rps3 proteins compared here. Open boxes represent sequences of less well conservea regions or sequence of no similarity between each other. X is the natural termination codon of *Chlamydomonas reinhardtii rps3* gene. The three domains (N-, S-, and C-domains of *Chlamydomonas reinhardtii* Rps3 protein) are shown below each fragment of the *rps3*; the number of amino acid residues they encode and the predicted sizes of polypeptides are indicated.

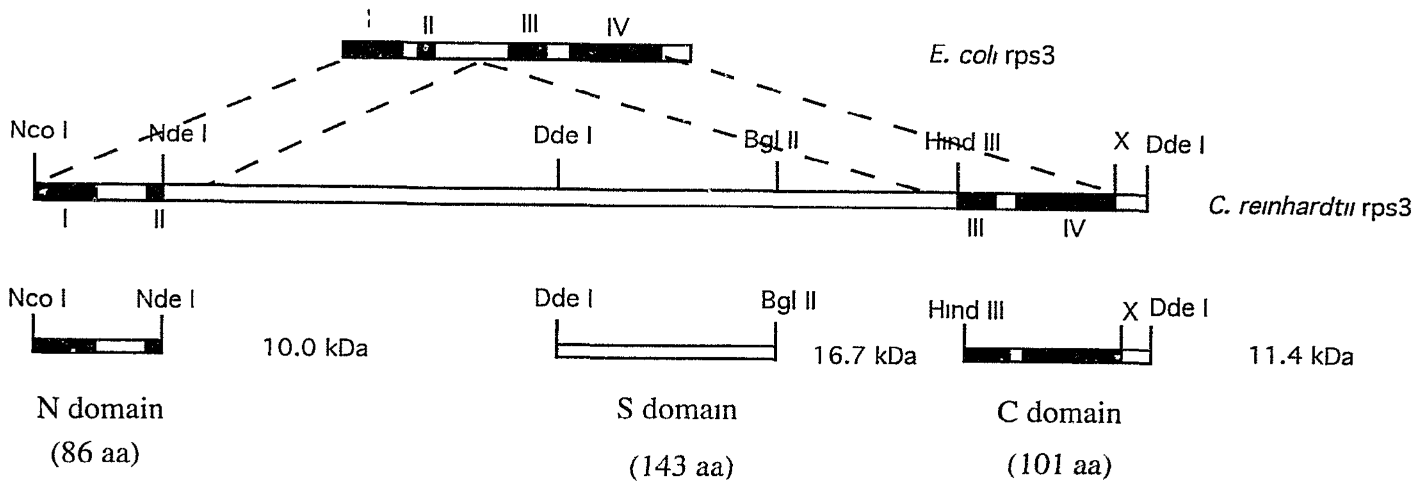

I
II
III
IV
E. coli rps3
Nco I
Nde I
Dde I
Bgl II
Hind III
X
Dde I
C. reinhardtii rps3
I
II
III
IV
Nco I
Nde I
10.0 kDa
Dde I
Bgl II
16.7 kDa
Hind III
X
Dde I
11.4 kDa
N domain
(86 aa)
S domain
(143 aa)
C domain
(101 aa)

Figure 5-12

residues (612 to 712) of the Rps3 protein, and contains the conserved regions III and IV found in conventional Rps3 proteins. These fragments were cloned into pMAL vector and proteins were expressed in *E. coli* and used to raise antibodies from rabbits as described in Materials and Methods section. Antibodies raised with these proteins were correspondingly designated as N-, S-, and C-domain specific antibodies.

In raising antibodies, proteins were expressed in *E. coli* from clones contain N-, S- and C-domains, purified by gel electroelution or amylose resin affinity column chromatography, and subsequently used as antigens. Antibodies were raised against each of the proteins (see Materials and Methods) and designated N-, S-, and C-domain specific antibodies. Due to the nature of antigen purification, antigens were actually only partially purified by either methods. For example, when the antigen was purified by affinity column, many minor protein bands were observed on an analytical SDS-PAGE gel stained with Coomassie blue; when it was purified by electroelution, any protein with the same apparent molecular weight would have co-purified. Either way, antigens would have contained a substantial amount of non-*rps3* related proteins. As a result, the corresponding anti-sera must have contained many non-specific antibodies. As expected, when anti-sera were used in Western blot analysis with *E. coli* total proteins as antigens, dozens of non-Rps3 protein bands were observed (data not shown).

b). Purification and specificity test of the domain-specific antibodies

To purify the domain-specific antibodies described above, the affinity purification by nitrocellulose membrane absorption method was used. However, non-specific antibodies against those proteins co-migrating with the antigen will be co-purified by this method. To solve this problem, two strips of membrane bearing the same part of the *rps3*-related protein but migrating at different positions were used in antibody purification. For example, N-specific antibodies were first purified from anti-sera with strip bearing protein expressed from clone pWC156 and then pWC129 (Figure 5-13), thus only the N-specific antibodies will be purified. The purified N-antibody was further purified with the membrane bearing protein expressed from clone pWC159 and was designated N1 antibody, as it would have specific activity to conserved sequence block I of a conventional Rps3 protein. The leftover antibody from the second round of purification was designated N2 antibody, as it would be specific to the conserved sequence block II of a conventional Rps3 protein.

The specificity of the purified antibodies was further tested with proteins expressed from clone pWC156, pWC159 and pWC160 in *E. coli.* As shown in Fig. 5-14, N1 antibody recognized proteins expressed from clone pWC156 (lane 3), pWC159 (lane 1), but not that from the clone pWC160 (lane 2), suggesting that it has the activity for the first 40 amino acid residues. N2 antibody only recognizes protein expressed from clone pWC156 (lane 6), not the protein from clone pWC159 (lane 4), suggesting that the activity against the first 40 amino acid residues has been completely depleted, and the remaining activity was directed against amino acid residues 41 to 86. Both antibodies recognized the protein expressed from clone pWC156 (lane 3 and lane 6), whereas neither of them recognized protein expressed from clone pWC160 (lane 2 and lane 5). The same strategy was also applied to the purification of S- and C-domain-specific antibodies, and the specificity of resulting antibodies was tested in a similar way (data not shown).

c). Western blot analysis of *rps3* gene products in *Chlamydomonas reinhardtii*

The expression of *rps3* gene in *Chlamydomonas reinhardtii* chloroplast was investigated by Western blot analyses with antibodies purified as described above. Previous evidence suggested that Rps3 protein is most likely an authentic ribosomal protein despite its very unusual gene structure, genome location, and an undetectable

Figure 5-13. Subclones used in antibody purification and specificity testing. Shown above is the schematic alignment of Rps3 protein from *C. reinhardtii* and from *E. coli*. pWC156, pWC159 and pWC160 are clones in pET vector, only the expressed protein parts are shown, and the relationship to the *C. reinhardtii* Rps3 protein is indicated. pWC129 is the N-domain cloned in pMAL vector, the protein expressed is a fusion protein, with the 42-kDa maltose-binding protein fused at its N-terminus as indicated (hatched box).

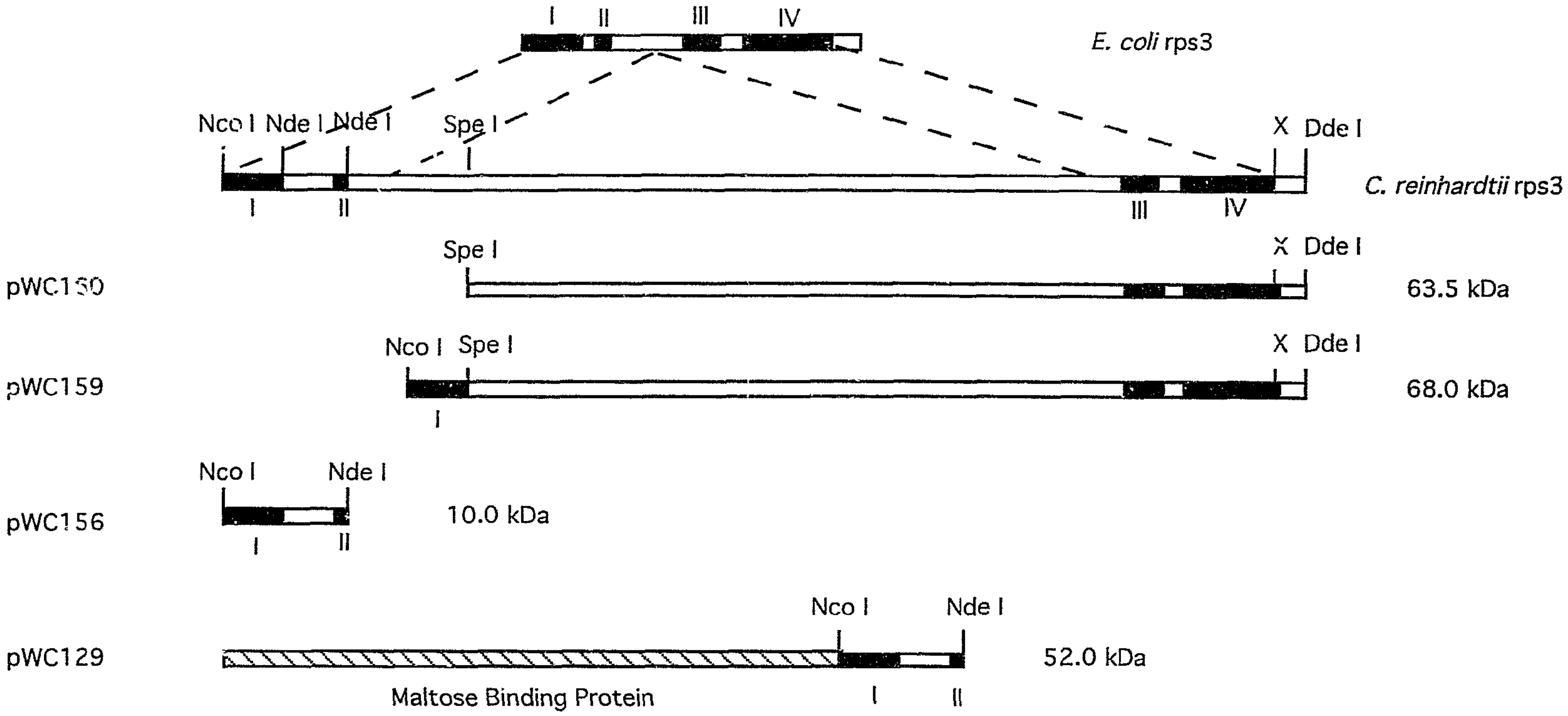

I
II
III
IV
E. coli rps3
Nco I
Nde I
Nde I
Spe I
X
Dde I
C. reinhardtii rps3
I
II
III
IV
Spe I
X
Dde I
pWC150
63.5 kDa
Nco I
Spe I
X
Dde I
pWC159
68.0 kDa
I
Nco I
Nde I
pWC156
10.0 kDa
I
II
Nco I
Nde I
pWC129
52.0 kDa
Maltose Binding Protein
I
II

Figure 5-13

Figure 5-14. The specificity of N1 and N2 antibodies. N antibody was first purified by adsorption to protein expressed from clone pWC129, and then purified against protein expressed from clone pWC156. At this point the antibody would contain activities to both conserved region I and region II of a conventional Rps3 protein. The antibody was then further purified by adsorption to protein expressed from clone pWC159; the resultant antibody was N1 antibody, which would contain activity to the conserved region I of a conventional Rps3 protein. The material remaining after the final purification was designated N2 antibody, which would contain activity to the conserved region II of the Rps3 protein. Lanes 1 and 4 contain proteins expressed from clone pWC159, lanes 2 and 5 contain proteins expressed from clone pWC160, lanes 3 and 6 contain proteins expressed from clone pWC156. Proteins are blotted onto a NitroPlus membrane after being resolved on SDS-PAGE, as described in the Materials and Methods section. Lanes 1 to 3 were blotted with N1 antibody, lanes 4 to 6 were blotted with N2 antibody.

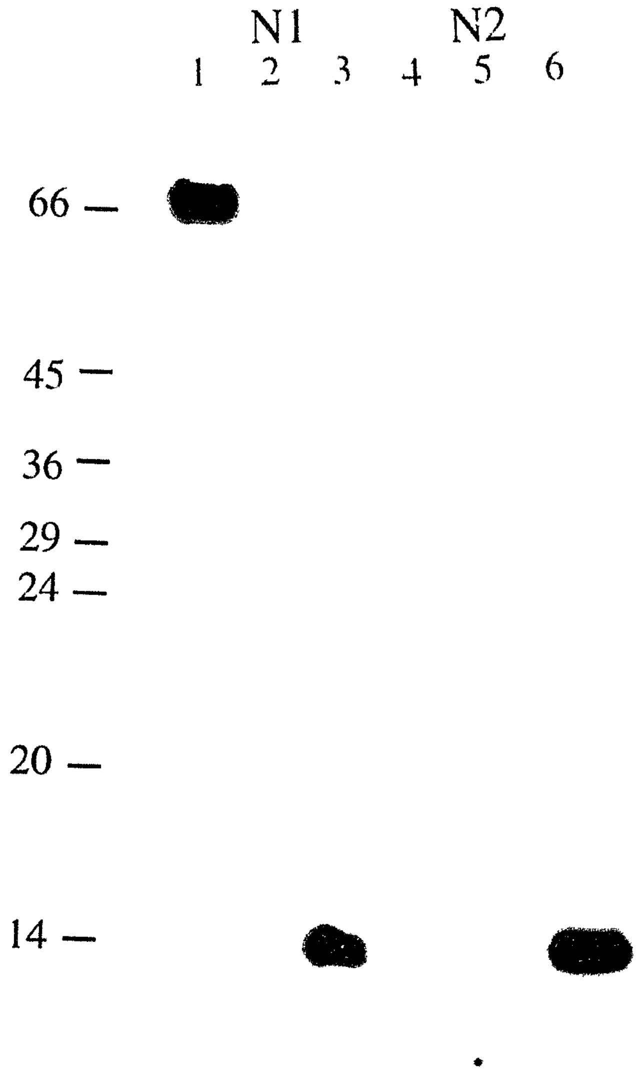


Figure 5-14

level of the corresponding mRNA in the steady-state RNA population (Liu et al. 1993). In Western blot, three fractions (the total cellular protein, the soluble protein, and the 30S small subunit of chloroplast ribosome) were examined for the presence of Rps3-related proteins. Typically, during the isolation of 30S ribosomal small subunit, a portion of the cell resuspension (4x 10^9 cells/ml in SA buffer, see Materials and Methods) was mixed with an equal volume of 2x protein loading buffer, heated in a boiling water bath for 5 min, centrifuged for 5 min, and the supernatant was taken as total cellular protein sample. After passing the cell resuspension through a FrenchPress and centrifugation, another portion was taken from the supernatant and boiled with equal volume of 2x protein loading buffer. This sample was used as the soluble protein fraction. The 30S ribosomal small subunit was isolated according to the protocol described by Schmidt et al. (1983) with modifications (see Materials and Methods); 5 µg of 30S ribosomal proteins were used in Western blot analysis. Protein samples from total cellular proteins, from soluble fraction, and from the isolated 30S fraction were resolved on analytical SDS-PAGE, blotted onto nitrocellulose (NitroPlus) membrane, and blotted with purified N1-, N2-, S-, and C-domain-specific antibodies. The amount of proteins loaded was adjusted so that the total cellular protein and soluble protein were prepared from the same number of cells, and the proteins loaded were close to the upper limit of the gel capacity. The 30S fraction is highly enriched for 30S ribosome small subunit, with 5 µg of 30S protein being prepared from 100 to 250 times the number of cells used for the total and soluble protein samples.

As shown in Fig. 5-15, N1 and N2 antibodies both recognized two protein bands in the 30S lane, of 83-kDa and 29-kDa (lanes 3 and 4), whereas S-domain-specific antibody recognized 83-kDa and 54-kDa protein bands in the 30S lane (lanes 5 and 6), suggesting that Rps3 is expressed as an 83-kDa precursor protein, then split into two parts, a 29-kDa protein consisting of the N-terminal part, and a 54-kDa protein containing of the S-domain part. S-domain-specific antibody did not recognize any protein band in the total and soluble protein fractions (data not shown). C-domain-specific antibody recognized three bands in the 30S fraction, of 83-kDa, 54-kDa, and 30-kDa (lane 7), but did not recognize any protein bands in the total cellular protein or in the soluble protein fraction (data not shown). The 83-kDa protein is recognized by all the antibodies, further suggesting that it is the Rps3 precursor protein. The 54-kDa protein is recognized by both S and C antibodies, suggesting that it is likely the Rps3 protein without the 29-kDa N-terminal part. The 29-kDa protein is recognized by both N1 and N2 antibodies, suggesting that it is the N-terminal part of the Rps3 precursor protein. C antibody also recognized a 30-kDa protein band, but this is not the same protein as the 29-kDa protein recognized by N1 and N2 antibodies, as demonstrated by two-dimensional gel electrophoresis followed by Western analyses with N1-, N2- and C-domain-specific antibodies (data not shown). It is likely it is just the C terminal part of the Rps3 protein. N1 antibody also recognized a 20-kDa protein band in the total cellular protein (lane 1) and in the soluble protein fraction (lane 2), but this band was not recognized by N2 antibody (data not shown), suggesting that it is unlikely to be the N-terminal portion of the Rps3 protein. The 20-kDa signal is strong in lane 1 but weak in lane 2, indicating that it is associated with cell membrane or other insoluble material. If it is the N-terminal part of the Rps3 protein, it should contain amino acid residues 41 to 86, and be recognized by N2 antibody. Overall, the results have shown that it is likely that the Rps3 protein is first expressed as a precursor and then cleaved into 29-kDa and 54-kDa proteins. Whether the 30-kDa protein is cleaved from the precursor or from the 54-kDa protein is unknown at this point.

Figure 5-15. The expression of *rps3* gene in *Chlamydomonas reinhardtii* chloroplast. Lane 1 is total cellular protein, lane 2 is the soluble protein fraction. They were prepared from the same number of cells (4 x 10^7 cells in each lane). Lanes 3 to 7 were 30S fractions, each lane containing 5 µg of 30S ribosomal proteins. Lanes 1 to 3 were blotted with N1 antibody, lane 4 was blotted with N2 antibody, lanes 5 and 6 were blotted with S-domain-specific antibody, lane 7 was blotted with C-domain-specific antibody. S- and C-domain-specific antibodies did not recognize any bands in the total cellular protein fraction and in the soluble protein fraction (data not shown). Lanes 1 to 5 were blotted from a preparative gel, lanes 6 to 7 were blotted from an analytical gel, but the amount of proteins loaded on lanes 3 to 7 were the same (5 µg). Protein markers for the preparative gel were a low-range marker from Sigma, those for the analytical gel were a middle-range marker from Promega. Sizes of protein markers are shown at the left side of gels.

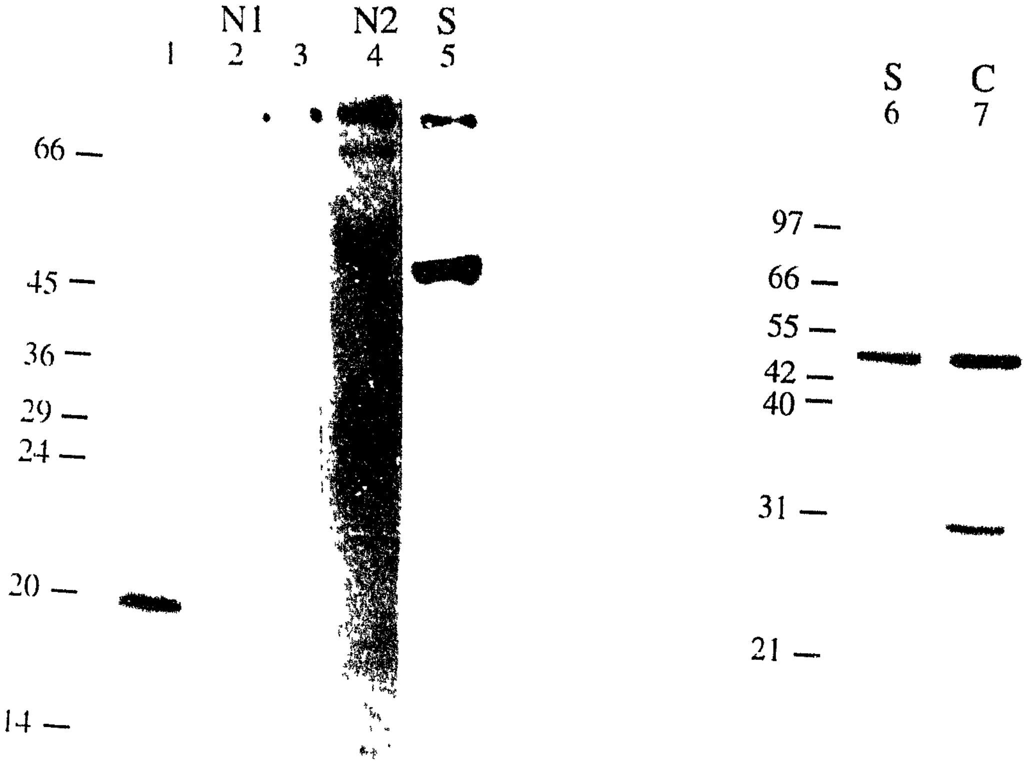


Figure 5-15

DISCUSSION

The structurally unusual *clpP* and *rps3* genes of *Chlamydomonas* chloroplasts have been investigated for their expression and protein products. These series of studies have identified the insertion sequence IS2 of ClpP as a degenerate intein whose protein splicing activity can be restored by a single amino acid substitution. These studies also revealed that the insertion sequence IS1 of ClpP is not removed from the precursor protein in *Chlamydomonas* under the conditions investigated, and the insertion sequence of Rps3 appears to result in a split of the precursor protein into several polypeptides. These conclusions and their implications are discussed below.

IS2 as a restorable degenerate intein

The modified IS2 sequence (IS2-m), containing a substitution of histidine for glycine near its C-terminus, clearly behaves as an active intein that is removed from a precursor protein through protein splicing. The central piece of evidence supporting this conclusion was the identification of a correctly spliced protein. When the IS2-m coding sequence was embedded in-frame with flanking coding sequences and expressed in *E. coli*, a 66-kDa spliced protein was produced in addition to the 117-kDa precursor protein. Initial indications of the 66-kDa protein as a spliced protein included its small and expected size and its peptide analysis profile. Micro-protein sequencing of the 66-kDa protein across the splice junction confirmed not only that the IS2-m sequence was precisely removed, but also that the two flanking sequences were joined by a bona fide peptide bond. These are hallmark characteristics of protein splicing.

The 66-kDa spliced protein is clearly a result of protein splicing, rather than RNA splicing or translational ribosome hopping. The IS2-m coding sequence lacks recognizable structural features of known introns. It also lacks potential hairpin-forming sequences that are required in know cases of ribosome hopping. Instead, it has strong features of a translated sequence, including being an open reading frame in-frame with its flanking sequences and having strong codon usage bias similar to other genes. More importantly, a spliced mRNA transcript was not detected even by using a very sensitive RT-PCR method, while an unspliced transcript was readily detectable. Another piece of evidence against RNA splicing or ribosome hopping came from the gene disruption experiment. When a termination codon was introduced into the IS2-m sequence, both the 117-kDa precursor protein and the 66-kDa spliced protein disappeared, while a 99-kDa truncated protein product accumulated as expected (Fig. 5-7). This indicated that translation through the IS2-m sequence is necessary for the production of the 66-kDa spliced protein. In contrast to RNA splicing and ribosome hopping, protein splicing requires the insertion sequence to be translated.

The protein splicing event should also produce an excised IS2-m polypeptide in addition to the correct spliced protein. The *E. coli* expression system used in identifying the spliced protein did not allow an identification of an excised IS2-m polypeptide. Unlike the spliced protein that can be affinity-purified on an amylose column, an excised IS2-m could not be easily purified from the *E. coli* cell lysate because of a lack of method for its purification and detection. An antibody against the IS2 polypeptide was not available. Under conditions used in the protein induction, an excised IS2-m polypeptide did not accumulate to a sufficiently high level to allow identification by Coomassie blue staining. However, later studies under improved conditions have detected the accumulation of a polypeptide with a size expected for an excised IS2-m polypeptide (Wu, personal communication). Furthermore, a protein corresponding to excised IS2-m in size was also detected in the wheat germ expression system. These observations and the clear identification of the spliced protein all point to the production of an excised IS2-m polypeptide in the protein splicing reaction, although a final proof may require IS2-m purification followed by micro-protein sequencing or Western blot analysis using IS2-specific antibodies.

In the *E. coli* expression system, a significant amount of the 117-kDa precursor protein was accumulated in addition to the spliced protein, indicating that the protein splicing reaction did not go to completion. This could be explained by a slow protein splicing reaction, but misfolding and precipitation of the precursor protein appeared to be important contributing factors. When the *E. coli* cell lysate was separated into soluble and insoluble fractions, a major fraction of the precursor protein was in the insoluble fraction, indicating the formation of insoluble inclusion bodies. When the protein induction was carried out at lower temperature (12 °C), little or no precursor protein was observed, although significant amount of the spliced protein was produced. Lower temperature is known to reduce the formation of insoluble inclusion bodies by lowering the rate of protein synthesis.

The protein splicing event observed with the IS2-m sequence is likely an autocatalytic reaction, suggesting that IS2-m is a self-splicing intein. Among all known inteins, only the intein associated with the DNA polymerase of *Pyrococcus* has been shown beyond doubt to be a self-splicing intein. A purified precursor protein containing this intein has been shown to undergo protein splicing *in vitro* without assistance of any other protein or RNA molecules. Inteins associated with the yeast VMA1 protein and the mycobacterial RecA protein were also suggested to be self-splicing inteins, because they support protein splicing in *E. coli* cells as well as in their native cells. The chloroplast-derived IS2-m intein sequence was shown to support protein splicing in two heterologous systems, namely in *E. coli* cells and in a wheat germ *in vitro* translation system. If a trans-acting protein or RNA molecule was required for the splicing, it would have to be present in both of these two very different cell systems.

It is not clear whether the IS2-m protein splicing follows the same pathway as determined for the *Pyroccocus* DNA polymerase intein, which involves nucleophilic attack, N-O or N-S shift, and formation of branched intermediate and succinimide ring. The IS2-m sequence contains all the necessary amino acid residues implicated in the above protein splicing pathway, suggesting that a similar pathway may be employed in the IS2-m protein splicing. However, the lack of overall sequence similarity between IS2-m and the *Pyroccocus* DNA polymerase intein may suggest differences between the two inteins in structure and function. In order for the protein splicing to take place, both the primary amino acid sequence and the tertiary structure are required. On one hand, some amino acid residues participate directly in the peptide cleavage and transpeptide reactions, and these residues would have to be conserved in the primary sequence. On the other hand, the tertiary structure may have to be formed in such a way that the two splicing junctions are brought into proximity with one another and the participating amino acid residues are placed in an appropriate orientation. While there is no easy way to know its tertiary structure, the primary sequence comparison did reveal that the ClpP IS2 has the molecular basis of an intein.

The IS2-m intein sequence likely contain all the structural information required for the protein splicing. This has been shown to be true for the yeast VMA1- and the *Pyrococcus* DNA polymerase-associated inteins, in which the intein was inserted into another unrelated gene and shown to undergo correct protein splicing, indicating that exteins (flanking sequences) do not contain information required for the protein splicing. Although the IS2-m intein has not been subjected to this type of test, its protein splicing activity clearly does not require an intact ClpP protein. The *E. coli* expression system produced a fusion protein consisting of the maltose-binding protein fused to the C-terminal portion of the ClpP protein, the wheat germ *in vitro* translation system produced a truncated ClpP protein consisting of only the C-terminal portion of the ClpP protein, and both proteins were shown to undergo protein splicing.

The IS2 intein sequence appears to have most, if not all, of the intein-associated structures recognized so far among known inteins. Its size (456 amino acid residues) is close to the size range (from 360 amino acid residues in Psp pol 2 to 538 amino acid residues in Tli pol 1) of other known inteins. Like all the other inteins, IS2 is bounded by a pair of nucleophilic residues, with a Cys residue at the N-terminus of IS2 and a Ser residue at the N-terminus of the downstream C-extein. The IS2 sequence also has seven sequence blocks (A to G) that appear to be somewhat conserved among known inteins. The sequence block A is present only when a intein sequence starts with Cys, not Ser, and the IS2 sequence starts with Cys and does have a sequence block A. In addition, the few universally conserved amino acid residues of known inteins were all found in IS2, except that the His residue near the C-terminus is replaced by a Gly residue.

On the other hand, overall sequence similarity between IS2 and other inteins is very low and practically undetectable. This is not surprising because overall sequence similarities among other inteins are also very low, unless the comparison is between two inteins of the same gene of two closely related species. Sequence similarities among different inteins are low even when the comparison is limited to only the seven conserved sequence blocks. For example, sequence identity of the seven sequence blocks is 21% between IS2 and the yeast Sce VMA intein, 21% between IS2 and the mycobacterial Mtu RecA intein, 31.5% between yeast Sce VMA intein and mycobacterial Mtu RecA intein (the two demonstrated splicing inteins starting with Cys), and 30% between mycobacterial Mtu RecA intein and Mle RecA intein. This low sequence similarity among different inteins may suggest that the inteins have independent origins, or they have undergone long periods of separate evolution, and functional constraint on the primary sequence of inteins may be very limited.

The unmodified (wildtype) IS2 is suggested to be a degenerate intein, because it does not support protein splicing either in *E. coli* cells or in the wheat germ translation system, although it can be restored to an active intein (IS2-m) by a single amino acid substitution. This degeneration of IS2 as an intein, if it indeed occurred in *Chlamydomonas eugametos*, would certainly include a critical His to Gly change near its C-terminus (in sequence block G), since this His residue is conserved in all other known inteins but replaced by Gly in IS2. In the yeast VMA1 intein, replacing this His residue by Gly completely abolishes its protein splicing activity, while replacement by Lys, Glu, Val, or Leu residues results in a lower rate of protein splicing (Cooper et al. 1993). It is also noted that a minimum of two nucleotide substitutions (C to G and A to G, one involving transition and the other involving transversion) are required to change a His codon (CAT) into a Gly codon (GGT), suggesting that substantial, but non-critical, additional mutations likely have occurred to other parts of the IS2 sequence.

It is not known whether the *Chlamydomonas eugametos* ClpP protein with its unmodified (wildtype) IS2 sequence can undergo protein splicing in the chloroplast. A ClpP protein in *Chlamydomonas eugametos* was not detected by using antibodies raised against *Chlamydomonas reinhardtii* ClpP, possibly due to several reasons. For example, the ClpP protein may be produced only under certain unknown conditions, or it may be accumulated at a very low level that escapes detection by Western blotting under the conditions used. Nevertheless, the fact that unmodified (wildtype) IS2 failed to support protein splicing in the two heterologous systems could suggest that IS2 is a degenerate (inactive) intein unable to splice *in vivo*. On the other hand, it can not be ruled out that the unmodified (wildtype) IS2 may supp rt protein splicing *in vivo* with assistance of yet unidentified trans-acting factors (protein factors or RNA editing, for instance). Also, an IS2 intein incapable of splicing would need to be tolerated by the ClpP protein structure and function, since the IS2-containing *clpP* gene has been described as essential for the organism.

The origin of IS2 intein remains an interesting but open question. IS2 was found only in the two interfertile species *Chlamydomonas eugametos* and *Chlamydomonas moewusii*. It is absent in all other *Chlamydomonas* species examined in this study, including *Chlamydomonas indica* which is closely related to *Chlamydomonas eugametos*. This suggests that IS2 intein was recently gained in the ClpP protein. The lack of sequence similarity between IS2 and the IS1 insertion sequence in the same protein suggests that the gain of IS2 is independent of the gain of IS1. A simple hypothesis is that the coding sequence of IS2 intein, as a mobile genetic element, "jumped" into the chloroplast *clpP* ger of the *Chlamydomonas eugametos* lineage from an unknown source, which could be another gene, another genome, or another organism. In this respect, it is interesting to note that IS2 contains two sequence motifs (sequence blocks C and E) similar to the LAGLI-DADG motifs characteristic of intron- and intein-associated endonucleases. Intron- and intein-associated endonuclease activities have been suggested, and sometimes demonstrated, to promote intron/intein mobility.

The identification of IS2 as a bona fide intein (although degenerate) has expanded the distribution of inteins to include a cell organelle gene, in addition to the previously identified intein-containing nuclear, archaebacterial, and eubacterial genes. In terms of host protein, ClpP protease represents the fourth protein found to contain an intein, with the previous three being an ATP se protein (VMA1), a DNA polymerase protein, and a DNA recombination protein (RecA). Although the number of known inteins is small, it has been suggested repeatedly that inteins may be more widespread and more numerous, partly because of a structural and evolutionary parallel between inteins and introns, and also because of the intein's characteristics as a mobile genetic element. The observation that IS2 represents a degenerate form of a self-splicing intein (IS2-m) raises the notion that inteins may exist in different forms or types. Like introns, inteins may turn out to include members that differ in structure and splicing mechanism, such as self-splicing, trans-splicing, and assisted splicing inteins. Inteins and protein splicing therefore present a new dimension in gene expression. The central dogma proposed nearly half a century ago predicted that RNA is an exact copy of information from one of the DNA strands, and that information presented in protein is co-linear with the information in the RNA molecule. This dogma has since been modified by discoveries of introns and RNA splicing, RNA editing, translational ribosome hopping (bypassing), as well as inteins and protein splicing.

The IS1-containing ClpP as a fusion protein

The insertion sequence IS1 of the *Chlamydomonas reinhardtii* chloroplast ClpP protein is not removed from the precursor protein either *in vivo* or *in vitro*, at least under conditions used in the present studies. Only a single polypeptide corresponding to the precursor protein was detected either by Western blotting analysis of *Chlamydomonas reinhardtii* total proteins or after expressing the *Chlamydomonas reinhardtii clpP* gene in *E. coli* cells. It is not known whether the IS1 sequence is removed from the precursor protein in *Chlamydomonas eugametos*, due to the inability to detect in this organism a ClpP protein of any form. It is very unlikely, however, that the *Chlamydomonas eugametos* IS1 would be removed from the precursor protein, because of its structural similarity to the IS1 of *Chlamydomonas reinhardtii*. Inside *Chlamydomonas*, it is still possible that the IS1 sequence may be removed from the precursor protein under certain unknown conditions. However, the fact that only the precursor protein was observed in logarithmically growing cells strongly suggests that a removal of IS1 is not required for normal cell growth. It is also possible, although very unlikely as discussed below, that the IS1 sequence could be restored to protein splicing by some unknown modifications.

The lack of protein splicing raises questions regarding the identity and evolutionary origin of the IS1 sequence. It is clearly an extra piece of sequence that was somehow inserted into the chloroplast ClpP protein in the *Chlamydomonas* lineage, because it is completely absent in ClpP proteins of all non-*Chlamydomonas* organisms studied so far. This insertion most likely occurred in a common ancestor of *Chlamydomonas*, since the IS1 sequence is present in all the *Chlamydomonas* species examined. The large size (286 amino acid residues) of IS1 distinguishes it from the short pieces (a few residues) of so called "gap sequences" frequently observed when comparing homologous sequences. One interesting possibility is that IS1, like IS2, is a degenerate intein. Unlike IS2 that has a more recent origin in the ClpP protein, IS1 clearly entered the ClpP protein at a much earlier time and may therefore have undergone a much longer period of degeneration. Interestingly, the IS1 sequence is still bounded by two nucleophilic (serine) residues, which is an important characteristic of known inteins. However, the IS1 sequence lacks recognizable sequence blocks that are conserved in known inteins including IS2, which could mean that such sequence blocks in IS1 have mutated beyond recognition, or that IS1 was a type of intein different from the others.

The observation that IS1 is not removed from the precursor protein by protein splicing has important implications for the structure and function of the chloroplast ClpP protein in *Chlamydomonas*. In *Chlamydomonas reinhardtii*, the presence of IS1 makes the precursor protein more than twice as large as a conventional ClpP protein. In *Chlamydomonas eugametos*, where both IS1 and IS2 are likely not to be removed from the precursor protein, the unprocessed protein would be five times as large as a conventional ClpP protein. It is likely that the flanking sequences (exteins) of IS1 and IS2 may fold together to form a conventional ClpP protein, with the IS1 and IS2 sequences looping out as spacer sequences. The presence of IS1 and IS2 sequences would have to be tolerated by the ClpP protein in its assembly and function. In *E. coli*, a ClpP protein needs to assemble into a multicomponent protease complex. Seven ClpP proteins form a disc-like ring structure, six ClpA proteins form another disc-like ring structure, and the functional protease complex contains two superimposed ClpP rings flanked at one or both ends by a ClpA ring (Kessel et al, 1995). Proteins to be degraded enter the protease complex at one end of the cylindrical structure and exit at the other end as short peptides or amino acids. Chloroplast ClpP proteases most likely assemble into a structure similar to that of *E. coli* ClpP protease, as indicated by a recent study in land plants. In the *Chlamydomonas* ClpP protease, the extra IS1 and IS2 sequences would be expected to locate on the outside surface of the cylindrical structure without preventing the structure and function of the protease complex.

It has been noted that the chloroplast ClpP protein appears to have a slower rate of evolution in *Chlamydomonas* than in plants. When the insertional sequences IS1 and IS2 are excluded, sequence identity between *C. eugametos* ClpP and *C. reinhardtii* ClpP is 95%, which is significantly greater than the 79% sequence identity between liverwort ClpP and ClpP from angiosperms. However, the evolutionary distance between these two *Chlamydomonas* species has been estimated to be greater than the distance between liverwort and angiosperms, based on the comparative analysis of chloroplast (Durocher et al, 1989, Turmel et al, 1993) and nuclear rRNA (Buchheim et al, 1990) sequences as well as various protein-coding sequences from both organelles. This discrepancy is unlikely to be a result of a recent horizontal transfer of *clpP* gene between these two *Chlamydomonas* species, since their IS1 sequences have diverged extensively. A more likely explanation is that the higher-than-expected sequence conservation is due to the presence of additional structural constraints, which could have been brought on by the presence of the IS1 and IS2 sequences.

It is also possible that the insertion sequence IS1 have acquired certain unknown cellular functions. Although this sequence has diverged extensively between *Chlamydomonas reinhardtii* and *Chlamydomonas eugametos*, it still maintains several highly conserved sequence blocks. They include one stretch of highly charged sequence located at each end of the IS1 sequence and several stretches of internal sequences. It is not possible at present to assign a likely function to these conserved sequence blocks, because they do not resemble any recognizable sequence motifs associated with known functions. On the other hand, if the IS1 sequence is simply a spacer sequence in the ClpP protein without any function, it would be difficult to explain why these conserved sequence blocks were maintained through the course of evolution. Furthermore, if the IS1 sequence is removed from the precursor protein under certain unidentified conditions, this may represent a regulatory step in the production and function of the ClpP protease.

The *rps3* gene in *Chlamydomonas reinhardtii*

The detection of Rps3-related protein products by Western blot suggests that the gene is expressed in the *Chlamydomonas reinhardtii* chloroplast. The *rps3* gene was first described as an *orf712* in *Chlamydomonas reinhardtii* chloroplast genome. Although all the four conserved domains of an Rps3 protein were present in the protein deduced from the *Chlamydomonas rps3* gene, there were several reasons to believe that it was more likely a pseudogene. First, it has an unusual structure in that the sequence similarity to a conventional Rps3 protein was only found at its N and C termini; two thirds of the sequence in the middle is not similar to any known protein. Second, it has an unusual location in the genome in that normally the gene for a conventional chloroplast-encoded Rps3 protein is located in a cluster of ribosomal protein genes equivalent to the *S10* operon in *E. coli*, however, the *rps3* gene in *Chlamydomonas reinnardtii* chloroplast genome is located more than 75 kbp away from the *S10*-like gene cluster. Third, it has an unusual expression pattern in that it lacks detectable mRNA by Northern blot (Fong et al, 1992). Gene disruption followed by chloroplast transformation demonstrated that this gene is essential to cell growth, with the mutant having a phenotype similar to that of other essential ribosomal protein gene mutations (Liu et al, 1993). Moreover, the conserved regions of a conventional Rps3 protein are present in the deduced protein sequence. All of these observations indicate that *rps3* is likely an authentic Rps3 protein gene. However, the unknown identity of the two thirds of the sequence from the middle of the *rps3* gene and the lack of evidence that this part of the sequence is spliced out at RNA level, raised interesting questions about how this gene is expressed and where the gene product(s) are functioning. The detection of Rps3-related protein products by high-quality domain-specific antibodies clearly suggested that the gene is expressed.

Although the quality of antibodies obtained was reasonably high, none of the antibodies except antibody N1 detected any protein band in a total cellular protein. If the *rps3* gene is indeed coding for an Rps3 protein, its products should be assembled into the small subunit of the chloroplast ribosome. Isolation of chloroplast ribosome small subunit from *Chlamydomonas reinhardtii* will enrich small subunit ribosomal proteins and make the rps3 gene products detectable by the antibodies. Furthermore, such an experiment should give an insight into how the gene is expressed, i.e., as just a precursor protein, a split protein, or a spliced protein. The isolation of chloroplast 30S ribosomal small subunit was based on a procedure described by Schmidt et al. (1983), and the procedure has proven successful in the same genetic system, *Chlamydomonas reinhardtii*. The identity of the isolated 30S fraction was further tested by Northern hybridization analysis with 16S rRNA-specific probes, and by Western blot analysis with antibodies known to be specific to *Chlamydomonas reinhardtii* chloroplast 30S ribosomal proteins (not Rps3, data not shown). The detection of Rps3-related protein products in the 30S ribosome fraction clearly indicates that *rps3* gene products are most likely functioning in the ribosome, and most likely functioning as Rps3 protein.

Multiple protein products are detected by high quality domain-specific antibodies in extensive Western blot analyses. Both N1 and N2 antibodies recognized an 83-kDa and a 29-kDa protein, S antibodies recognized an 83-kDa and a 54-kDa protein, and C antibodies recognized an 83-kDa, a 54-kDa and a 30-kDa protein. The 83-kDa protein is most likely the precursor protein as it is recognized by all the antibodies, and the size of the protein is the same as that expected from the *rps3* gene. As the 29-kDa protein is recognized by both N1 and N2 antibodies, the 54-kDa protein is recognized by both S and C antibodies, and 29-kDa plus 54-kDa is 83-kDa, it is likely that the precursor protein is expressed and then cleaved to give rise to the 29-kDa and 54-kDa proteins. C antibodies also recognized a 30-kDa protein, which is not recognized by any other antibodies, suggesting that it only contains the C part of the precursor. As all of the Rps3 protein products are detected in the 30S ribosomal subunit, including the precursor, it is likely that the processing takes place after the assembly of precursor into the ribosome.

The 20-kDa protein recognized by N1 antibody is likely a non-specific protein, for the following reasons. First, it is not in the 30S ribosomal small subunit fraction, but is associated with membrane or other insoluble materials. Second, if the 20-kDa protein is expressed from the *rps3* gene, it should contain approximately 175 amino acid residues from the N-terminus, and be recognized by both N1 and N2 antibodies. Third, if the 20-kDa protein composes the first 40 amino acid residues plus other portions of the precursor protein, it should be recognized by either S or C antibodies. Based on this argument, the 20-kDa component is likely a non-specific protein. Since the N terminus of Rps3 protein is highly conserved in species from *E. coli* to chloroplasts of high plants, but the sequence similarity is limited only to the first 40 amino acid residues, it is also possible that the 20-kDa protein is Rps3 protein from other source, such as mitochondria.

The 30-kDa protein detected by C antibody is unlikely to be the product of an internal translation initiation of the *rps3* gene transcript, as there is no methionine (Met) residue at an appropriate position. The closest Met residues are at positions 400 or 472, which would result in a 36-kDa or a 27-kDa protein if the internal translation initiation started from one of these two Met residues. It is more likely that the protein derives from the cleavage of the precursor or from the 54-kDa protein. The absence of the other part of the protein after cleavage could be due to the rapid degradation or loose association to the 30S ribosomal subunit, so that it is lost during the isolation of the 30S subunit. In conclusion, at least four Rps3-related proteins exist in *Chlamydomonas reinhardtii* chloroplast. It is unclear though, whether pieces of Rps3-related protein products are being held together and functioning as conventional Rps3 protein, or if only some of the pieces are functional after cleavage. One interesting possibility is that the 29-kDa N-terminal and the 30-kDa C-terminal polypeptides associate to form a conventional Rps3 protein, the middle part being cut away and quickly degraded, and with the 54-kDa protein serving as the intermediate.

The origin of extra sequence found in *Chlamydomonas* Rps3 proteins remains an interesting but open question. When homologous Rps3 protein sequences are aligned, sequence similarity is usually higher at N- and C-termini, while one-third of the sequence in the middle is relatively low in sequence similarity (see Chapter III, for example), suggesting that the sequences at both ends of an Rps3 protein are functionally more conserved than that of the middle part. It is interesting to note that *rps3* encoded in maize

(Hunt and Newton, 1991) and rapeseed (Ye et al, 1993) mitochondrial genomes are also larger that t eir counterpart in *E. coli*, and the expansion also seems to be in the middle of the gene, however, the middle portions are much smaller and not similar to that in the *Chlamydomonas* homolog.

When two-dimensional gel electrophoresis profiles of chloroplast ribosomal proteins are compared to those of *E. coli* ribosomes, extra spots are normally observed (Subramanian, 1993), indicating that chloroplast ribosomes contain more proteins than *E. coli* ribosomes do. So far, at least five of these extra ribosomal proteins have been characterized, all of which are found to be encoded in the nucleus. The results of this study add another possibility to the origin of these extra ribosomal proteins, that is, some of the "new" proteins may be encoded in the chloroplast genome and multi-protein products may be produced by cleavage of a protein precursor encoded by a single gene in the chloroplast genome.

VI. REFERENCES

Baldauf, S.L. and Palmer, J.D. 1990. Evolutionary transfer of the chloroplast tufA gene to the nucleus. Nature (London). 344:262-265

Barettino, D., Feigenbutz, M., Valcárcel, R., Stunnenberg, H.G. 1993. Improved method foı PCR-mediated mutagenesis. Nucleic Acids Research. 22:541-542.

Belfort, M. 1990. Phage T4 introns: self-splicing and mobility. Annual Review of Genetics. 24:363-385.

Bell-Pedersen, D., Quirk, S., Clyman, J., and Belfort, M. 1990. Intron mobility in phage T4 is independent upon a distinctive class of endonucleases and independent of DNA sequences encoding the intron core: mechanistic and evolutionary implications. Nucleic Acids Research. 18:3763-3770.

Bonen, L., Cunningham, R.S., Gray, M.W., Doolittle, W.F. 1977. Wheat embryo mitochondrial 18S ribosomal RNA: evidence for its prokaryotic nature. Nucleic Acids Research. 4:663-671.

Boudreau, E., Otis, C., Turmel, M. 1994. Conserved gene clusters in the highly rearranged chloroplast genomes of *Chlamydomonas moewusii* and *Chlamydomonas reinhardtii.* Plant Molecular Biology. 24:585-602.

Boynton, J.E., Gillham, N.W., Lambowitz, A.M. 1980. Ribosomes. (Chamblis, G., Craven, G.R., Davies, J., Davis, K., Kahan, L., Nomura, M. ads., University Park Press). 903-950.

Briat, J., Letoffe, S., Mache, R., Rouviere-Yaniv, J. 1984. Similarity between the bacterial histone-like protein HU and a protein from spinach chloroplasts. FEBS Letters. 172:75-79.

Brinkmann, H., Martinez, P., Quigley, F., Martin, W., Cerff, R. 1987. Endosymbiotic origin and codon bias of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. Journal of Molecular Evolution. 26:320-328.

Brown, T.A., Waring, R.B., Scazzocchio, C., Davies, R.W. 1985. The *Aspergillus nidulans* mitochondrial genome. Current Genetics. 9:113-117.

Cech, T.R. 1990. Self-splicing of group I introns. Annual review of Biochemistry. 59:543-568

Cedergren, R., Gray, M.W., Abel, Y., Sankoff, D. 1988. The Evolutionary relationships among known life forms. Journal of molecular evolution. 28:98-112.

Christine, B. 1990. The Cyanelle *S10 spc* Ribosomal protein gene operon from *Cyanophora paradoxa*. Molecular General Genetics. 224:221-231

Christopher, D.A. and Hallick R.B. 1989. *Euglena gracilis* chloroplast ribosomal protein operon: a new chloroplast gene for ribosomal protein L5 and description of a novel organelle intron category designated group III. Nucleic Acids Research. 17:7591-7608.

Colleaux, L., Michel-Wolwertz, M.R., Matagne, R.F., Dujon, B. 1990. The apocytochrome b gene of *Chlamydomonas smithii* contains a mobile intron related to both *Saccharomyces* and *Neurospora* introns. Molecular General Genetics. 223:288-96.

Cooper, A.A., Chen, Y-J., Lindorfer, M.A. and Stevens, T.H. 1993. Protein splicing of the yeast *TFP1* intervening protein sequence: a model for self-excision. The EMBO Journal. 12:2575-2583.

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Research. 16:10881-10890.

Cote, M.J. and Turmel, M. 1995. In vitro self-splicing reactions of chloroplast and mitochondrial group-I introns in *Chlamydomonas eugametos* and *Chlamydomonas moewusii.* Current Genetics. 27:177-183.

Cote, V., Mercier, J.P., Lemieux, C., Turmel, M. 1993. The single group I intron in the chloroplast *rrnL* gene of *Chlamydomonas humicola* encodes a site-specific DNA endonuclease (I-ChuI). Gene. 129:69-76.

Covello, P.S. and Gray, M.W. 1992. Silent mitochondrial and active nuclear genes for subunit 2 of cytochrome c oxidase (cox2) in soybean: evidence for RNA-mediated gene transfer. The EMBO Journal 11:3815-3820.

Cozens, A.L., Walker, J.E., Phillips, A.L., Huttly, A.K., Gray, J.C. 1986. A sixth subunit of ATP synthase, an F_0 component, is encoded in the pea chloroplast genome. EMBO Journal. 5:217-222.

Davis, E.O., Jenner, P.J., Brooks, P.C., Colston, M.J., and Sedgwick, S.G. 1992. Protein splicing in the maturation of *M. tuberculosis* RecA protein: a mechanism for tolerating a novel class of intervening sequence. Cell. 71:201-210.

Davis, E.O., Sedgwick, S.G., ana Colston, M.J. 1991. Novel structure of the *recA* locus of *Mycobacterium tuberculosis* implies processing of the gene product. Journal of Bacteriology. 173:5653-5662.

Davis, E.O., Thangaraj, H.S., Brooks, P.C. and Colston, M.J. 1994. Evidence of selection for protein introns in the RecAs of pathogenic mycobacteria. The EMBO Journal. 13:699-703.

Douglas, S.E. 1991. Unusual organization of a ribosomal protein operon in the plastid genome of *Guillardia theta*: evolutionary considerations. Current Genetics. 19:289-294.

Douglas, S.E. 1992 A SecY homologue is found in the plastid genome of *Guillardia theta*. FEBS Letter. 298:93-96.

Douglas, S.E., Murphy, C.A., Spencer, D.F., and Gray, M.W. 1991 Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. Nature. 350:148-151.

Douglas, S.E. and Turner, S. 1991. Molecular evidence for the origin of plastids from a cyanobacterium-like ancestor. Journal of Molecular Evolution. 33:267-273.

Drlica, K. and Rouviere, Y.J. 1987. Histone-like proteins of bacteria. Microbiology Review. 51:301-319.

Dujon, B. 1989. Group I introns as mobile genetic elements: facts and mechanistic speculations — a review. Gene. 82:91-114.

Durrenberger, F. and Rochaix, J.D. 1991. Chloroplast ribosomeal intron of *Chlamydomonas reinhardtii*: *in vitro* self-splicing, DNA endonuclease activity and *in vivo* mobility. EMBO Journal. 10:3495-3501.

Eschbach, S., Hofmann, C.J.B., Maier, U-G., Sitte, P., Hansmann, P. 1991. A eukaryotic genome of 660 kb: electrophoretic karyotype of nucleomorph and cell nucleus of the cryptomonad alga, *Pyrenomonas salina*. Nucleic Acids Research. 19:1779-1781.

Fersht, A. 1977. Enzyme structure and machenism. 303-315.

Flashner, Y. and Gralla, J. 1988. DNA Dynamic flexibility and protein recognition: differential stimulation by bacterial histone-like protein Hu. Cell. 54:713-721.

Fong, S.E. and Surzycki, S.J. 1992. Organization and structure of plastome *psbF*, *psbL*, *petG* and *ORF712* genes in *Chlamydomonas reinhardtii*. Current Genetics. 21:527-530.

Froehlich, J.E., Poorman, R., Reardon, E., Barnum, S.R., Jaworski, J.G. 1990. Purification and characterization of acyl carrier protein from two cyanobacteria species. European Journal of Biochemistry. 193:817-825.

Gantt, J.S. 1988. Nucleotide sequences of cDNAs encoding four complete nuclear-encoded plastid ribosomal proteins. Current Genetics. 14:519-528.

Gantt, J.S., Baldauf, S.L., Calie, P.J., Weeden, N.F., Palmer, J.D. 1991. Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. The EMBO Journal. 10:3073-3078.

Gauthier, A., Turmel, M., Lemieux, C. 1991. A group I intron in the chloroplast large subunit rRNA gene of *Chlamydomonas eugametos* encodes a double-strand endonuclease that cleaves the homing site of this intron. Current Genetics. 19:43-47.

Gibbs, S.P. 1981. The chloroplast endoplasmic reticulum: structure, function, and evolutionary significance. International Review of Cytology. 72:49-99.

Gimble, F.S. and Thorner, J. 1992. Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. Nature. 357:301-306.

Giovannoni, S.J., Turner, S., Olson, G.J., Barns, S., Lane, D.J., Pace N.R. 1988. Evolutionary relationships among cyanobacteria and green chloroplasts. Journal of Bacteriology. 170:3584-3592.

Goguel, V., Dalahodde, A., and Jacq, C. 1992. Connections between RNA splicing and DNA intron mobility in yeast mitochondria: RNA maturase and DNA endonuclease switching experiments. Molecular Cell Biology. 12(2):696-705.

Gottesman, S., Squires C., Pichersky, E., Carrington, M., Hobbs, M., Mattick, J.S., Dalrymple, B., Kuramitsu, H., Shiroza, T., Foster, T., Clarke W.P., Ross, B., Squires C.L., Maurizi, M.R. 1990. Conservation of the regulatory subunit for the Clp ATP-dependent protease in prokaryotes and eukaryotes. Proc. Natl. Acad. Sci. USA. 87:3513-3517.

Gray, J.C., Hird, S.M., Dyer, T.A. 1990. Nucleotide sequence of a wheat chloroplast gene encoding the proteolytic subunit of an ATP-dependent protease. Plant Molecular Biology. 15:947-950.

Gray, M.W. 1988. Organelle origins and ribosomal RNA. Biochemistry and Cell Biology. 66:325-348.

Gray, M.W. 1992. The Endosymbiont hypothesis revisited. International Review of Cytology. 141:233-357.

Gray, M.W. and Doolittle, W.F. 1982. Has the endosymbiont hypothesis been proven? Microbiological Reviews. 46:1-42.

Gray, M.W., Sankoff, D., Cedergren, R.J. 1984. On the Evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. Nucleic Acids Research. 12:5837-5852.

Greenwood, A.D. 1974. The cryptophyta in relation to phylogeny and photosynthesis. Electron Microscopy (Sanders, J.V. and Goodchild, D.J. eds, Canberra: Australian Academy of Sciences). 566-567.

Gu, H.H., Xu, J., Gallagher, M., and Dean, G.E. 1993. Peptide splicing in the vacuolar ATPase subunit A from *Candida tropicalis*. The Journal of Biological Chemistry. 268:7372-7381.

Gutell, R.R., Schnare, M.N., Gray, M.W. 1990. A compilation of large subunit (23S-like) ribosomal RNA sequences presented in a secondary structure format. Nucleic Acids Research. 18 (Suppl.):2319-2330.

Higgins, D.J., Bleasby, A.J., Fuchs, R. 1992. Improved software for multiple sequence alignment. Cabios. 8:189 191.

Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., Kondo, C., Honji, Y., Sun, C-R., Meng, B-Y., Li, Y- Q., Kanno, A., Nishizawa, Y., Hirai, A., Shinozaki, K., Sugiura, M. 1989. The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct rRNA genes accounts for a major plastid DNA inversion during the evolution of cereals. Molecular General Genetics. 217:185-194.

Hirata, R. and Anraku, Y. 1992. Mutations at the putative junction sites of the yeast VMA1 protein, the catalytic subunit of the vacuolar membrane H^+-ATPase, inhibit its processing by protein splicing. Biochemistry and Biophysics Research Communication. 188:40-47.

Hodges, R.A., Perler, F.B., Noren, C.J. and Jack, W.E. 1992. Protein splicing removes intervening sequences in an archaea DNA polymerase. Nucleic Acids Research. 20:6153-6157.

Huang, W.M., Ao, S-Z, Casjens, S., Orlandi, R., Zeikus, R., Weiss, R., Winge D., Fang, M. 1988. A persistent untranslated sequence within bacteriophage T4 DNA topoisomerase gene 60. Science. 239:1005-1012.

Huang, C., Wang, S., Chen, L., Lemieux, C., Otis, C., Turmel, M., Liu, X-Q. 1994. The *Chlamydomonas* chloroplast clpP gene contains translated large insertion sequences and is essential for cell growth. Molecular General Genetics. 244:151-159.

Hunt, M.D. and Newton, K.J. 1991. The NCS3 mutation: genetic evidence for the expression of ribosomal protein genes in *Zea mays*. The EMBO Journal. 10:1045-1052.

Jacquier, A and Dujon, B. 1985. An intron-encoded protein is active in a gene conversion process that spreads an intron into a mitochondrial gene. Cell. 41:383-394.

Jahn, O., Hartmann, R.K., Erdmann, V.A. 1991. Analysis of the *spc* ribosomal protein operon of *Thermus aquaticus*. European Journal of Biochemistry. 197:733-740.

Jaworski, J.G., Post, B.M., Ohlrogge, J.B. 1989. Site-directed mutagenesis of the spinach acyl carrier protein-I prosthetic group attachment site. European Journal of Biochemistry. 184:603-609.

Johnson, C.H., Kruft, V., Subramanian, A.R. 1990. Identification of a plastid-specific ribosomal protein in the 30S subunit of chloroplast ribosomes and isolation of the cDNA clone encoding its cytoplasmic precursor. Journal of Biological Chemistry. 265:12790-12795.

Johnson, C.H. and Subramanian, A.R. 1991. Chloroplast ribosomal protein *L15*, like *L1, L13* and *L21*, is significantly larger than its *E. coli* homologue. FEBS. 282:268-272.

Kane, P.M., Yamashiro, C.T., Wolczyk, D.F., Neff, N., Goebl, M., Stevens, T.H. 1990. Protein splicing converts the yeast *TFP1* gene product to the 69-kDa subunit of the vacuolar H+-adenosine triphosphatase. Science. 250:651-657.

Katayama-Fujimura, Y., Gottesman, S., Maurizi, M.R. 1987. A multiple-component, ATP-dependent protease from *Escherichia coli*. Journal of Biological Chemistry. 262:4477-4485

Kessel, M., Maurizi, M.R., Kim, B., Kocsis, E., Trus, B.L., Singh, S.K., and Steven, A.C. 1995. Homology in structural organization between *E. coli ClpP* protease and the eukaryotic 26S proteasome. Journal of Molecular Biology. 250:587-594.

Kowallik, K.V., Stoebe, B., Schaffran, I., Kroth-Pancic, P., Freier, U. 1995. The chloroplast genome of a chlorophyll a+c-containing alga, *Odontella sinensis*. Plant Molecular Biology Reporter. 13:336-342.

Kroh, H.E., Simon, L.D. 1991. The ClpP Component of Clp protease is the sigma-32 dependent heat shock protein F21.5. Journal of Bacteriology. 172:6026-6034.

Kuo, T.M., Ohlrogge, J.B. 1984. The Primary structure of spinach acyl carrier protein. Archive of Biochemistry and Biophysics. 234:290-296.

Lambowitz, A.M. 1989. Infectious introns. Cell 56:323-326.

Liaud, M-F., Zhang, D-X., and Cerff, R. 1990. Different Intron loss and endosymbiotic transfer of chloroplast glyceraldehyde-3-phosphate dehydrogenase genes to the nucleus. Proc. Natl. Acad. Sci. USA. 87:8918-8922.

Liu, X-Q., Huang, C., Xu, H. 1993. The unusual *Rps3*-like *Orf712* is functionally essential and structurally conserved in *Chlamydomonas*. FEBS. 336:225-230.

Ludwig, M. and Gibbs, S.P. 1985. DNA is present in the nucleomorph of cryptomonads: further evidence that the chloroplast evolved from a eukaryotic endosymbiont. Protoplasma. 127:9-20.

Ma, D.P., King, Y.T., Kim, Y., Luckett, W.S. 1992. The group I intron of apocytochrome b gene from *Chlamydomonas smithii* encodes a site-specific endonuclease. Plant Molecular Biology. 18:1001-1004.

Marshall, P. and Lemieux, C. 1991. Cleavage pattern of the homing endonuclease encoded by the fifth intron in the chloroplast large subunit rRNA-encoding gene of *Chlamydomonas eugametos*. Gene. 104:241-245.

Martin, W. and Cerff, R. 1986. Prokaryotic features of a nucleus-encoded enzyme. European Journal of Biochemistry. 159:323-331.

Maurizi, M.R., Clark, W.P., Katayama, Y., Rudikoff, S., Pumphrey, J., Bowers, B., Gottesman, S. 1990a. Sequence and structure of ClpP, the proteolytic component of the ATP-dependent Clp protease of *Escherichia coli*. Journal of Biological Chemistry. 265:12536-12545.

Maurizi, M.R., Clark, W.P., Kim, S-H., Gottesman, S. 1990b. ClpP represents a unique family of serine proteases. Journal of Biological Chemistry. 265:12546- 12552.

McFadden, G.I. 1990. Evidence that cryptomonad chloroplasts evolved from photosynthetic eukaryotic endosymbionts. Journal of Cell Science. 95:303-308.

McFadden, G.I., Gilson, P.R., Douglas, S.E. 1994a. The photosynthetic endosymbiont in cryptomonad cells produces both chloroplast and cytoplasmic-type ribosomes. Journal of Cell Science. 107:649-657.

McFadden, G.I., Gilson, P.R., Hofmann, C.J., Adcock, G.J., Maier, U-G. 1994b. Evidence that an amoeba acquired a chloroplast by retaining part of an engulfed eukaryotic alga. Proc. Natl. Acad. Sci. USA. 91:3690-3694.

Michalowski, C.B., Pfanzagl, B., Löffelhardt, W., Bohnert, H.J. 1990. The cyanelle *S10 spc* ribosomal protein gene operon from *Cyanophora paradoxa*. Molecular General Genetics. 224:222-231.

Michel, F., and Dujon, B. 1986. ibid. 46:323-

Moore, T., Keegstra, K. 1993. Characterization of a cDNA clone encoding a chloroplast-targeted Clp homologue. Plant Molecular Biology. 21:525-537.

Mota, E.M., and Collins, R.A. 1988. Independent evolution of structural and coding regions in a *Neurospora* mitochondrial intron. Nature 332:654-656.

Nomura, M., Yates, J.L., Dean, D., Post, L. 1980. Feedback regulation of ribosomal protein gene expression in *Escherichia coli*: structural homology of ribosomal RNA and ribosomal protein mRNA. Proc. Natl. Acad. Sci. USA. 77:7084-7088.

Nomura, M., Gourse, R., and Baughman, G. 1984. Regulation of the synthesis of ribosomes and ribosomal components. Annual Review of Biochemistry. 53:75-117.

Oda, K., Yamato, K., Ohta, E., Nakamura, Y., Takemura, M., Nozato, N., Akashi, K., Kanegae, T., Ogura, Y., Kohchi, T., Ohyama, T. 1992. Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA. Journal of Molecular Biology. 223:1-7.

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, A., Inokuchi, H., Ozeki, H. 1986. Chloroplast gene organization deduced from complete sequence of liverword *Marchantia polymorpha* chloroplast DNA. Nature. 322:572-574.

Palenik, B. and Haselkorn, R. 1992. Multiple evolutionary origins of prochlorophytes, the chlorophyll b-containing prokaryotes. Nature (London). 355:265-267.

Palmer, J.D. 1985. Comparative organization of chloroplast genomes. Annual Review of Genetics. 19:325-354.

Perler, F.B., Comb, D.G., Jack, W.E., Moran, L.S., Qia.. , B., Kucera R.B., Benner. J., Slatko, B.E., Nwankwo, D.O., Hempstead, S.K., Carlow, C.K.S., and Jannasch, H. 1992. Intervening Sequences in an Archaea DNA Polymerase Gene. Proc. Natl. Acad. Sci. USA. 89:5577-5581.

Perler, F.B., Davis, E.O., Dean, G.E., Gimble, F.S., Jack, W.E., Neff, N., Noren, C.J., Thorner, J., Belford, M. 1994. Protein splicing elements: inteins and exteins--a definition of terms and recommended nomenclature. Nucleic Acids Research. 22:1125-1127

Perlman, P.S., and Butow, R.A. 1989. Mobile introns and intron-encoded proteins. Science. 246:1106-1109.

Pettijohn, D. 1988. Histone-like proteins and bacterial chromosome structure. Journal of Biological Chemistry. 263:12793-12796.

Pietrokovski, S. 1994. Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. Protein Science. 3:2340-2350.

Post-Beittenmiller, M.A., Hlousek-Radojcic, A., Ohlrogge, J.B. 1989. DNA sequence of a genomic clone encoding an *Arabidopsis* acyl carrier protein (ACP). Nucleic Acids Research. 17:1777-

Queen, C. and Corn, L.J. 1984. A comprehensive sequence analysis program for the IBM personal computer. Nucleic Acids Research. 12:581-599.

Recipon, H.. Perasso, R., Adoutte, A., Quetier, F. 1992. ATP synthase subunit c/III/9 gene sequences as a tool for interkingdom and metaphytes molecular phylogenies. Journal of Molecular Evolution. 34:292-303.

Reith, M. 1995. Molecular biology of rhodophyte and chromophyte plastids. Annual Review of Plant Physiology and Plant Molecular Biology. 46:549-575.

Reith, M. and Munholland, J. 1993. A high-resolution gene map of the chloroplast genome of the red alga *Porphyra purpurea*. Plant Cell. 5:465-475.

Robertson, D., Boynton, J.E., Gillham, N.W. 1990. Cotranscription of the wild-type chloroplast *atpE* gene encoding the CF1/CFo epsilon subunit with the 3' half of the *rps7* gene in *Chlamydomonas reinhardtii* and characterization of frameshift mutations in *atpE*. Molecular General Genetics. 221:155-163.

Rose, R.E., DeJesus, C.E., Moylan, S.L., Ridge, N.P., Scherer, D.E., Knauf, V.C. 1987. The nucleotide sequence of a cDNA clone encoding acyl carrier protein (ACP) from *Brassica campestris* seeds. Nucleic Acids Research. 15:7197-

Raff, R.A. and Mahler, H.R. 1972. The non symbiotic origin of mitochondria. Science. 177:575-582

Schmid, M.B. 1990. More than Just "histone-like" proteins. Cell. 63:451-453.

Schmidt, J., Srinivasa, B.R., Weglöhner, W., Subramanian, A.R. 1993. A small novel chloroplast ribosomal protein (S31) that has no apparent counterpart in the *E. coli* ribosome. Biochemistry and Molecular Biology International. 29:25-32.

Schmidt, K.M. and Ohlrogge, J.B. 1990. A root acyl carrier protein-II from spinach is also expressed in leaves and seeds. Plant Molecular Biology. 15:765-778.

Schmidt, R.J., Richardson, C.B., Gillham, N.W., Boynton, J.E. 1983. Sites of synthesis of chloroplast ribosomal proteins in *Chlamydomonas*. the Journal of Cell Biology. 96:1451-1463.

Schnare, M.N. and Gray, M.W. 1982. 3'-terminal sequence of wheat mitochondrial 18S ribosomal RNA: further evidence of a eubacterial evolutionary origin. Nuclei Acids Research. 10:3921-3932.

Schwartz, R.M. and Dayhoff, M.O. 1981. Chloroplast origins: inferences from protein and nucleic acid sequences. Annals of the New York Academy of Sciences. 361:260-269.

Sharma, M., Ellis, R.L., and Hinton, D.M. 1992. Identification of a family of bacteriophage T4 genes encoding proteins similar to those present in group I introns of fungi and phage. Proc. Natl. Acad. Sci. USA. 89:6658-6662.

Shih, M-C., Lazar, G., Goodman, H.M. 1986. Evidence in favor of the symbiotic origin of chloroplasts: primary structure and evolution of tobacco glyceraldehyde-3-phosphate dehydrogenases. Cell. 47:73-80.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, BY., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H., Sugiura, M. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 5:2043-2049.

Shub, D.A., and Goodrich-Blair, H. 1992. Protein introns: A new home for endonucleases. Cell. 71:183-186.

Spencer, D.F., Schnare, M.N. Gray, M.W. 1984. Pronounced structural similarities between the small subunit ribosomal RNA genes of wheat mitochondria and *Escherichia coli*. Proc. Natl. Acad. Sci. USA 81:493-497.

Squires, C., Squires, C.L. 1992. The Clp proteins: proteolysis regulators or molecular chaperones? Journal of Bacteriology. 174: 1081-1085.

Subramanian, A.R. 1993. Molecular genetics of chloroplast ribosomal proteins. Trends in Biochemical Science. 18:177-181.

Subramanian, A.R., Stahl, D., Prombona, A. 1991. The molecular biology of plastids. (Bogorad, L. and Vasil, I.K., ads. Academic Press). 191-21⌐.

Sugiura, M. 1992. The chloroplast genome. Plant Molecular Biology. 19:149-168.

Szostak, J., Orr-Weaver, T.L., Rothstein, R.J. 1983. ibid. 33:25-

Tanaka, I., Appelt, K., Dijk, J., White, S., Wilson, K. 1984. 3-Å resolution structure of a protein with histone-like properties in prokaryotes. Nature. 310:376-381.

Thompson, A.J., Yuan, X., Kudlicki, W., Herrin, D.L. 1992. Cleavage and recognition pattern of a double-strand-specific endonuclease (I-creI) encoded by the chloroplast 23S rRNA intron of *Chlamydomonas reinhardtii.* Gene. 119:247-251.

Turmel, M., Cote, V., Otis, C., Mercier, J.P., Gray, M.W., Lonergan, K.M., Lemieux, C. 1995a. Evolutionary transfer of ORF-containing group I introns between different subcellular compartments (chloroplast and mitochondrion). Molecular Biology and Evolution. 12:533-545.

Turmel, M., Mercier, J.P., Cote, V., Otis, C., Lemieux, C. 1995b. The site-specific DNA endonuclease encoded by a group I intron in the *Chlamydomonas pallidostigmatica* chloroplast small subunit rRNA gene introduces a single-strand break at low concentrations of Mg^{2+}. Nucleic Acids Research. 23:2519-25.

Turmel, M., Otis, C. 1994. The Chloroplast gene cluster containing *psbF, psbL, petG* and *rps3* is conserved in *Chlamydomonas.* Current Genetics. 27:54-61.

Turner, S., Burger-Wiersma, T., Giovannoni, S.J., Mur, L.R., Pace, N.R. 1989. The relationship of a prochlorophytes *Prochlorothrix hollandica* to green chloroplasts. Nature (London). 337:380-382.

Umezono, K. and Ozeki, H. 1987. Chloroplast gene organization in plants. Trends in Genetics. 3:281-287.

van der Boogaart, P., Samallo, J., Agsteribbe, E. 1982. Similar genes for a mitochondrial ATPase subunit in the nuclear and mitochondrial genomes of *Neurospora crassa.* Nature (London). 298:187-189.

Vanaman, T.C., Wakil, S.J., Hill, R.L. 1968. The complete amino acid sequence of the acyl carrier protein of *Escherichia coli*. Journal of Biological Chemistry. 243:6420-6431.

Wahleithner, J.A. and Wolstenholme, D.R. 1988. Ribosomal protein S14 genes in broad bean mitochondrial DNA. Nucleic Acid Research. 16:6897-6913.

Wallace, C.J.A. 1993. The curious case of protein splicing: mechanistic insights suggested by protein semisynthesis. Protein Science. 2:697-705.

Wang, S. and Liu, X-Q. 1991. The plastid genome of *Guillardia theta* encodes an *Hsp70*-like protein, a histone-like protein, and an acyl carrier protein. Proc. Natl. Acad. Sci. USA. 88:10783-10787.

Watkins, B.A., Davis, A.E., Cocchi, F., Reitz, M.S., Jr. 1993. A rapid method for site-specific mutagenesis using larger plasmids as templates. BioTechniques. 15:700-704.

Weglöhner, W. and Subramanian, A.R. 1993. Nucleotide sequence of maize chloroplast *rpl32*: completing the apparent set of plastid ribosomal protein genes and their tentative operon organization. Plant Molecular Biology. 21:543-548.

Wittmann, H.G. 1982. Components of bacterial ribosomes. Annual Review of Biochemistry. 51:155-183.

Woese, C.R., Stackebrandt, E., Weisburg, W.G., Paster, B.J., Madigan, M.T., Fowler, V.J., Hahn, C.M., Blanz, P., Gupta, R., Nealson, K.H., Fox, G.E. 1984. Systematic and applied Microbiology. 5:315-326.

Woodson, S.A. and Cech, T.R. 1989. Reverse self-splicing of the tetrahymena group I intron: implication for the directionality of splicing and for intron transposition. Cell. 57:335-345.

Xu, M-Q., Comb, D.G., Paulus, H., Noren, C.J., /shao, Y., and Perler, F B. 1994. Protein splicing: an analysis of the branched intermediate and its resolution by succinimide formation. EMBO Journal. 13:5517-5522.

Xu, M-Q., Southworth, M.W., Mersha, F.B., Hornstra, L.J., and Perler, F.B. 1993. In vitro protein splicing of purified precursor and the identification of a branched intermediate. Cell. 75:1371-1377.

Yamada, E.W., Dotzlaw, H., Huzel, N.J. 1991. Isolation of histone-like proteins from mitochondria of bovine heart. Preparative Biochemistry. 21:11-23.

Ye, F., Bernhardt, J.. Abel, W.O. 1993. Genes for ribosomal proteins *S3, L16, L5* and *S14* are clustered in the mitochondrial genome of *Brassica napus L.* Current Genetics. 24:323-329.

Zhou, D.X. and Mache, R. 1989. Presence in the stroma of chloroplasts of a large pool of a ribosomal protein not structurally related to any *Escherichia coli* ribosomal protein. Molecular General Genetics. 219:204-208

Zinn, A.R. and Butow, R.A. 1985. Non-reciprocal exchange between alleles of the yeast mitochondrial 21S rRNA gene: kinetics and the involvement of a double-strand break. Cell. 40:887-895.

Zurawski, G., and Zurawski, S.M. 1985. Structure of the *Escherichia coli S10* ribosomal protein operon. Nucleic Acids Research. 13:4521-4526.