A Study of Two Cyanobacterial Inteins

by

Hong Wu

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at

Dalhousie University

Halifax, Nova Scotia, Canada

July, 1998

0-612-36595-6

DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "A Study of Two Cyanobacterial Inteins"

by Hong Wu

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

Dated: September 2, 1998

External Examiner

Research Supervisor

Examining Committe

[redacted]

DALHOUSIE UNIVERSITY

DATE September 12, 1998

AUTHOR Hong Wu

TITLE A study of two cyanobacterial inteins

DEPARTMENT OR SCHOOL Biochemistry

Degree PhD Convercation Fall Year 1998

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

[redacted]

Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEFE EXERTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.

For my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Inteins are protein sequences embedded in-frame within precursor protein sequences and excised during a maturation process termed protein splicing. Protein splicing is a post-translational event involving precise excision of the intein sequence and concomitant ligation of the flanking sequences (N- and C-exteins) by a normal peptide bond.

The *dnaB* gene of cyanobacterium *Synechocystis* sp. strain PCC6803 was shown to encode a DNA helicase that has a 429-aa intervening sequence. This intervening sequence was identified as an intein, based on sequence analysis and a demonstration of protein splicing with or without the native exteins when tested in *E. coli* cells. A centrally located 275 aa sequence (residues 107-381) of this intein was deleted without loss of the protein splicing activity, resulting in a functional mini-intein of 154 aa. This mini-intein was then split into two separate fragments: a 106-aa N-terminal fragment containing intein motifs A and B, and a 48-aa C-terminal fragment containing intein motifs F and G. These two intein fragments, when produced in *E. coli* cells, carried out efficient protein *trans*-splicing. These results indicate that the N- and C-terminal regions of the *Ssp* DnaB intein, whether covalently linked with each other or not, can come together through non-covalent interactions to form a protein splicing domain that is functionally sufficient and structurally independent from the centrally located endonuclease domain of the intein. Mutagenesis of the 154-aa mini-intein revealed some internal residues that are involved in the protein splicing activity of this intein.

A naturally occurring split intein, capable of protein *trans*-splicing, was identified in a DnaE protein of the cyanobacterium *Synechocystis* sp. strain PCC6803. The N- and C-terminal halves of DnaE (catalytic subunit α of DNA polymerase III) are encoded by two separate genes, *dnaE-n* and *dnaE-c*, respectively. These two genes are located 745,226 bp apart in the genome and on opposite DNA strands. The *dnaE-n* product consists of a N-extein sequence followed by a 123-aa intein sequence, while the *dnaE-c* product consists of a 36-aa intein sequence followed by a C-extein sequence. The N- and C-extein sequences together reconstitute a complete DnaE sequence, which is interrupted by the intein sequences inside the domain that interacts with the β and τ subunits of DNA polymerase III. The two intein sequences together reconstitute a split mini-intein that not only has intein-like sequence features but also exhibited protein *trans*-splicing activity when tested in *E. coli* cells.

LIST OF ABBREVIATIONS

A, absorbance

aa, amino acid

°C, degree Celsius

ddH$_2$O, distilled, deionized water

dNTP, 2'-deoxynucleoside-5'-triphosphate

DTT, dithiothreitol

EDTA, ethylenediaminetetraacetate

EtBr, ethidium bromide

FPLC, fast protein liquid chromatography

g, gravity

HPLC, high performance liquid chromatography

h, hour

IPTG, isopropyl-β-D-thiogalactopyranoside

kb, kilobase

kbp, kilobase pair

kDa, kiloDalton

µg, microgram

µl, microliter

µM, micromolar

min, minute

ml, milliliter

mM, millimolar

M, molar

ng, nanogram

ORF, open reading frame

PCR, polymerase chain reaction

PVDF, polyvinylidene difluoride

rpm, revolutions per minute

SDS, sodium dodecyl sulfate

SDS-PAGE, SDS polyacrylamide gel electrophoresis

sec, second

TEMED, N, N, N', N'-tetramethylenediamine

Tris, tris(hydroxymethyl)-aminomethane

UV, ultraviolet

ACKNOWLEDGMENTS

I take this opportunity to thank those people who have helped me, in one way or another, during the completion of the research described in this thesis.

I would like to thank my supervisor, Dr. Paul X-Q. Liu for giving me the opportunity to do research work in his laboratory. I am particularly appreciative of the guidance, encouragement and patience that Dr. Liu has displayed throughout the course of my study. Gratitude is also extended to the members of this lab, both past and present, for their help and friendship. Special thanks are due to Ms. Zhuma Hu for her invaluable technical assistance and friendship.

I would like to express my thanks to my supervisory committee members, Dr. C. J. A. Wallace, Dr. R. A. Singer, and Dr. W. F. Doolittle, for their insightful discussions, suggestions and guidance as my research progressed.

During the course of this study, I enjoyed productive collaborations with the research groups at New England Biolabs, Inc. I would like to thank Dr. M. Q. Xu, for his supervision and help during my stay at NEB. My gratitude is also extended to the other members of Dr. Xu's lab for their support and friendship.

My appreciation also goes to my family members. I thank my husband Ping He for his constant support during my study. I greatly appreciate the encouragement and support from my parents. Without their unselfish sacrifice, I could hardly finish this work. And, thanks to my son, Kevin, for the inspiration and motivation he gives to me.

Chapter I Literature Review

1.1 Definition of intein and protein splicing:

An intein is a protein sequence embedded in-frame within a precursor protein sequence and excised during a maturation process termed protein splicing (Perler *et al*, 1994: Perler, 1998). Protein splicing is a recently discovered post-translational modification process. It involves the precise excision of an intervening fragment (the **intein**) from a precursor protein and the concomitant ligation of the flanking sequences (N- and C-**exteins**) by a normal peptide bond (Cooper and Stevens, 1995). Unlike introns, an intein coding sequence is not removed from RNA transcripts, but is translated in-frame as a part of the protein in which it is inserted. The intein sequence is then removed at the protein level to give two protein products: the mature host protein and the excised intein (Fig. 1-1). Protein splicing therefore adds another layer of complexity to the Central Dogma of molecular biology, because it changes the co-linearity between a gene and its protein product. Many inteins are also bi-functional elements, harboring an endonuclease activity that mediates the spread of inteins in addition to the protein splicing activity that catalyzes the intein's excision from the precursor protein. Both of the identified functions are for the intein preservation and dissemination, whereas no apparent benefit to the host protein or the organism has been found. Inteins are therefore believed to be selfish elements (Pietrokovski, 1998).

1.2 Distribution of inteins:

The first reported intein is encoded in the nuclear *VMA1* gene of *Saccharomyces cerevisiae* (Kane *et al.*, 1990; Hirata *et al.*, 1990). This gene encodes the 69-kDa catalytic subunit of the vacuolar H^+-ATPase. In *S. cerevisiae*, the sequence of the *VMA1* gene is interrupted by an in-frame insertion of a coding sequence for a 50-kDa protein that lacks homology to any known ATPase subunit but exhibits 31% sequence identity to the yeast

Figure 1-1 Schematic illustration of intein and protein splicing, and their contrast to intron and RNA splicing.

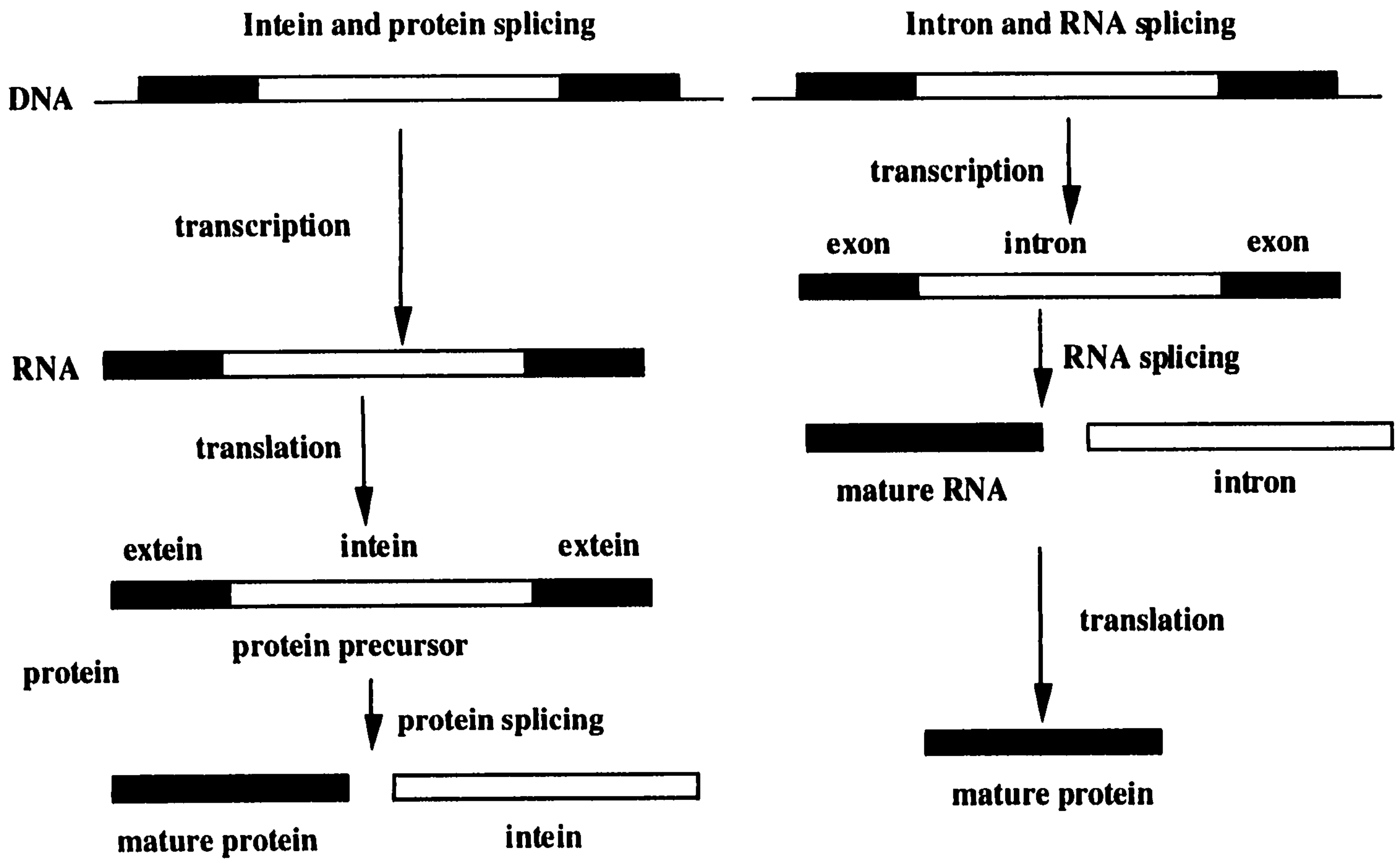

Intein and protein splicing
DNA
transcription
RNA
translation
extein
intein
extein
protein
protein precursor
protein splicing
mature protein
intein
Intron and RNA splicing
transcription
exon
intron
exon
RNA splicing
mature RNA
intron
translation
mature protein

Figure 1-1

HO endonuclease (Cooper *et al.*, 1993; Kane *et al.* 1990). The continuous open reading frame predicted a protein of 119 kDa, with the N- and C-exteins exhibiting 73% and 77% amino acid sequence identity to the equivalent V-ATPase subunit of *Neurospora crassa* (Kane *et al.*, 1990; Hirata *et al.*, 1990). A precise deletion of the intervening sequence produced a *VMA1* allele that fully complemented a *vma1* null allele and produced a V-ATPase subunit of the predicted size. This result indicated that the intein is not present in the final protein product. Analysis of the RNA sequence of this gene dic' not reveal any recognizable characteristics of known RNA introns. Northern blot analysis detected only a single mRNA of the length for translation of a 119-kDa precursor protein. This finding excluded the possibility of RNA splicing. Introduction of a stop codon in the intervening sequence showed that the continuity of the open reading frame is necessary for the production of the 69-kDa mature protein. Kinetic studies also showed that the 69-kDa mature protein and the 50-kDa intein were produced at equal rate, consistent with the synthesis of a precursor protein which is posttranslationally processed into two proteins. When the protein was made *in vitro* and in *E. coli*, two proteins (a 69-kDa protein and a 50-kDa protein) were generated. Since the process can take place in yeast, *E. coli*. and a rabbit reticulocyte lysate, it is likely to be an auto-catalytic reaction.

To date, more than fifty inteins and putative intein sequences have been described. A complete list of inteins can be found at http://www.neb.com/neb/inteins.html. Some of the inteins were identified experimentally, but most of them were predicted from DNA sequence. Several short conserved sequence motifs found in the known inteins helped in identifying new inteins. A large number of putative inteins were identified in *Methanococcus jannaschii* due to the genome sequencing project and the use of these short conserved sequence motifs (Bult *et al.*, 1996). On the other hand, the finding of many new putative inteins also helped to refine these conserved motifs (Perler *et al.*, 1997; Pietrokovski, 1998). Now a combination of four criteria is used to identify new inteins: (i) an in-frame insertion in a gene that has a previously sequenced homologue lacking the insertion; (ii) the presence of intein motifs C and E, which are also found in homing endonucleases (they are also called dodecapeptide motifs or LAGLIDADG motifs); (iii) the presence of other conserved intein motifs; (iv) the presence of Ser, Thr or Cys at the intein N-terminus, His-Asn at the intein C-terminus and Ser, Thr or Cys following the downstream splice site.

Inteins have been found in organisms spanning all three domains of life (eukaryote, eubacteria, and archaebacteria, see Table 1-1). Among the four inteins found in eukaryotes, two are encoded in chloroplast genes (*Ceu* ClpP and *Ppu* DnaB), while the other two are encoded in nuclear genes (*Sce* VMA and *Ctr* VMA). Seventeen inteins are found in eubacteria. Most of them are from *Mycobacterium* (10) and *Synechocystis* sp. PCC6803 (4). Among the 33 inteins found in archaebacteria, 19 of them are from 14 *Methanococcus jannaschii* genes.

The host proteins in which inteins are found are quite divergent. Many of the inteins are located in proteins that are involved in DNA replication, DNA repair, transcription, or translation (Table 1-2). Some of them are located in metabolic enzymes such as phosphoenolpyruvate synthase, anaerobic ribonucleotide triphosphate reductase, UDP-glucose dehydrogenase, ClpP protease, vacuolar ATPase and glutamine-fructose 6-phosphate transaminase. The known inteins have sizes ranging from 335 to 548 aa. There are four exceptions: the DnaB intein in *Porphyra purpurea* (150 aa) (Reith and Munholland, 1995), the GyrA intein in *Mycobacterium xenopi* (198 aa) (Perler *et al.*, 1997), the RIR1 intein in *Methanobacterium thermoautotrophicum* strain ΔH (134 aa) (Pietrokovski, 1998), and the KlbA intein in *Methanococcus jannaschii* (168 aa) (Dalgaard *et al.*, 1997; Pietrokovski, 1998). These small inteins lack the centrally located endonuclease domain

Table 1-1 Number of inteins found in different organisms

Organism	Number of inteins
Eukaryote	
Saccharomyces cerevisiae (nuclear)	1
Candida tropicalis (nuclear)	1
Chlamydomonas eugametos (chloroplast)	1
Porphyra purpurea (chloroplast)	1
Eubacteria	
Mycobacterium flavescens	2
Mycobacterium gordonae	1
Mycobacterium kansasii	1
Mycobacterium leprae	3
Mycobacterium malmoense	1
Mycobacterium xenopi	1
Mycobacterium tuberculosis	1
Rhodothermus marinus	1
Synechocystis sp. PCC6803	4
Bacillus subtilis SP beta phage	1
Deinococcus radiodurans	1
Archaebacteria	
Methanococcus jannaschii	19
Methanobacterium thermoautotrophium (delta H strain)	1

Table 1-1 Number of inteins found in different organisms (continued)

Organism	Number of inteins
Pyrococcus furiosus	3
Pyrococcus spp. GB-D	1
Pyrococcus spp. KOD	2
Thermococcus spp. TY	3
Thermococcus litoralis	2
Thermococcus fumicolans	2

Table 1-2 Number of inteins found in different host proteins

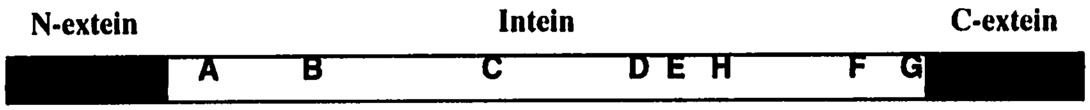
Host protein	Number of inteins
DNA polymerases	13
DNA gyrase A subunit	6
DNA gyrase B subunit	1
Reverse gyrase	1
DNA topoisomerase I	1
DNA helicase (DnaB)	4
DNA replication factor C (DnaX)	4
DNA recombinase (RecA)	3
RNA polymerase subunit	2
Transcription factor	1
Translation initiation factor	1
rNTP reductase	2
Ribonucleotide-diphosphate reductase	4
UDP-glucose dehydrogenase	1
Vacuolar ATPase subunit	2
ClpP protease	1
Phosphoenolpyruvate synthase	1
Glutamine-fructose-6P transaminase	1
KlbA protein	1
unidentified proteins	3

(including the two dodecapeptide motifs). They may represent naturally occurring minimal inteins.

1.3 Conserved intein motifs:

Sequence identities among inteins are generally very low, except for the presence of several conserved short sequences (Fig. 1-2). These short sequences include the highly conserved nucleophilic residue (cysteine, serine or threonine) at the N terminus of the intein and at the beginning of the C-terminal extein. In most inteins, the C terminus of the intein consists of His-Asn preceded by a stretch of hydrophobic residues. In addition, the inteins usually carry two dodecapeptide motifs, which are characteristic of homing endonucleases (Lambowitz and Belfort, 1993). Since even the most conserved regions of inteins constitute a rather minimal sequence alignment, it has been difficult to identify new inteins based only on the conserved intein motifs. By using methods for detecting weak, conserved sequence features to analyze the limited number of intein sequences, Pietrokovski proposed a more reliable method for identifying new inteins (Pietrokovski, 1994). Seven conserved sequence motifs (motifs A-G) in the inteins covering the three most conserved regions (the N-extein splice junction, the C-extein splice junction and the two dodecapeptide motifs) were proposed as criteria for intein identification. The conserved motif A is at the N terminus of the intein and contains the chemically essential Ser or Cys. Motif B contains a polar residue (most often Thr) three amino acids prior to a conserved His in all inteins. Motifs C and E are the dodecapeptide motifs required for endonuclease activity. Motif D is characterized by a conserved basic amino acid (usually Lys) and a Pro residue. Motif F contains aromatic residues on both sides of several acidic and hydrophobic residues. Motif G defines the C-terminal splice junction, and it contains the three conserved C-terminal junction residues (His-Asn-Cys/Ser/Thr) preceded by four hydrophobic residues. A new motif, motif H, was later defined between motif E and F (Perler *et al.*, 1997). It is characterized by one or more Ser or Thr residues in position 1-3,

Figure 1-2 Conserved intein motifs. Top: schematic illustration of conserved intein motifs. The consensus sequences (Perler *et al.*, 1997) and example sequences are shown below. Symbols in the consensus sequences: h, hydrophobic residues (V, L, I, A, M); a, acidic residues (D, E); r, aromatic residues (F, Y, W); p, polar residues (S, T, C); ., non-conserved residues; *, gap introduced into motif F; underlined residues, conserved in all inteins; capital letters, single letter amino acid code.



	motif A	motif B	motif C	motif D	motif E	motif H	motif F	motif G
consensus	Ch..Dp.hhh..G	G..h.hT..H.hhh	LhG..hhaG	.K.IP..h	.L.GhFahDG	p.S..hh..h..LL..hGI	rVYDLpV**a..HNFh	NGhhhHN
Sce VMA	CFAKGTNVLMADG	LLKFTCNATHELVV	LLGLWIGDG	VKNIPSFL	FLAGLIDSDG	TIHTSVRDGLVSLARSLGL	YGITLS**DDSDHQFL	NQVVVHN
Mxe GyrA	CITGDALVALPEG	GLRVTGTANHPLLC					PVYSLRV*DTADHAFL	NGFVSHN
Ceu ClpP	CLTSDHTVLTTRG	GVDLFVTPNHRMYV	FFGLWIANG	NKYLPDWV	LLNSLCLGNC	STSERFANDVSRLALHAGT	PVYCLTG***PNNVFY	KAVWTGN
Ppu DnaB	CISKFSHIMWSHV	EKYLELTSNHKILT					NVFDFAA**NPIPNFL	NNIIVHN
Rma DnaB	CLAGDTLITLADG	GRSIRATANHRFLT	LLGHLIGDG	EKKVPALL	FLRHLWATDG	TTSSYQLARDVQSLLLRLG	EVFDLTV**PGPHNFV	NDIIAHN
Ssp DnaB	CISGDSLISLAST	GRTIKATANHRFLT	LLGHLIGDG	EKFVPNQV	FLRHLWSTDG	TSSEKLAKDVQSLLLKLGI	EVFDLTV**PGPHNFV	NDIIVHN

Figure 1-2

a central hydrophobic region containing several Leu, and a Gly followed by a hydrophobic residue.

1.4 Mechanism of protein splicing:

Soon after the discovery of the first intein, several mechanisms for protein splicing were proposed. In any proposed mechanism of protein splicing, cleavage of the two peptide bonds in the precursor protein and the formation of a new peptide bond between N- and C-exteins are the two aspects that must be included. A commonly accepted mechanism was proposed by Xu and coworkers at New England Biolabs, Inc. (Xu and Perler, 1996; Chong *et al.*, 1996). In this model, the protein splicing pathway involves four steps (Fig. 1-3). The first step is an N-O/N-S acyl shift at the first nucleophilic residue of the intein. The acyl rearrangement moves the N-extein to the side chain of the first residue (Ser, Thr or Cys) of the intein. The invariant His in motif B may facilitate this acyl shift at the N-terminal splice junction. A transesterification occurs when the ester bond is attacked by the hydroxyl group of the first nucleophilic residue of the C-extein, resulting in transfer of the N-extein to the side chain of the first residue of the C-extein, forming a branched intermediate. The branched intermediate is then resolved by cyclization of the invariant intein C-terminal Asn to form a succinimide ring. Asn cyclization cleaves the peptide bond between the intein and the C-extein. The highly conserved His preceding the Asn helps in Asn cyclization. Succinimide formation leaves the N-extein attached to the side chain of the first residue of the C-extein via an ester bond, which is resolved by a spontaneous O-N acyl rearrangement resulting in the N-extein being linked to the C-extein via a native peptide bond.

This model was based on and evolved from some earlier models. The first model for a mechanism of protein splicing was proposed by Wallace (Wallace, 1993). In his model, Wallace suggested that the splicing reaction is initiated by N-O/N-S acyl shifts at

Figure 1-3 Mechanism of protein splicing. The mechanism involves four steps: the formation of an ester intermediate, the formation of a branched intermediate, excision of the intein, and spontaneous O/N acyl rearrangement to form the ligated exteins. Adapted from Xu and Perler, 1996.

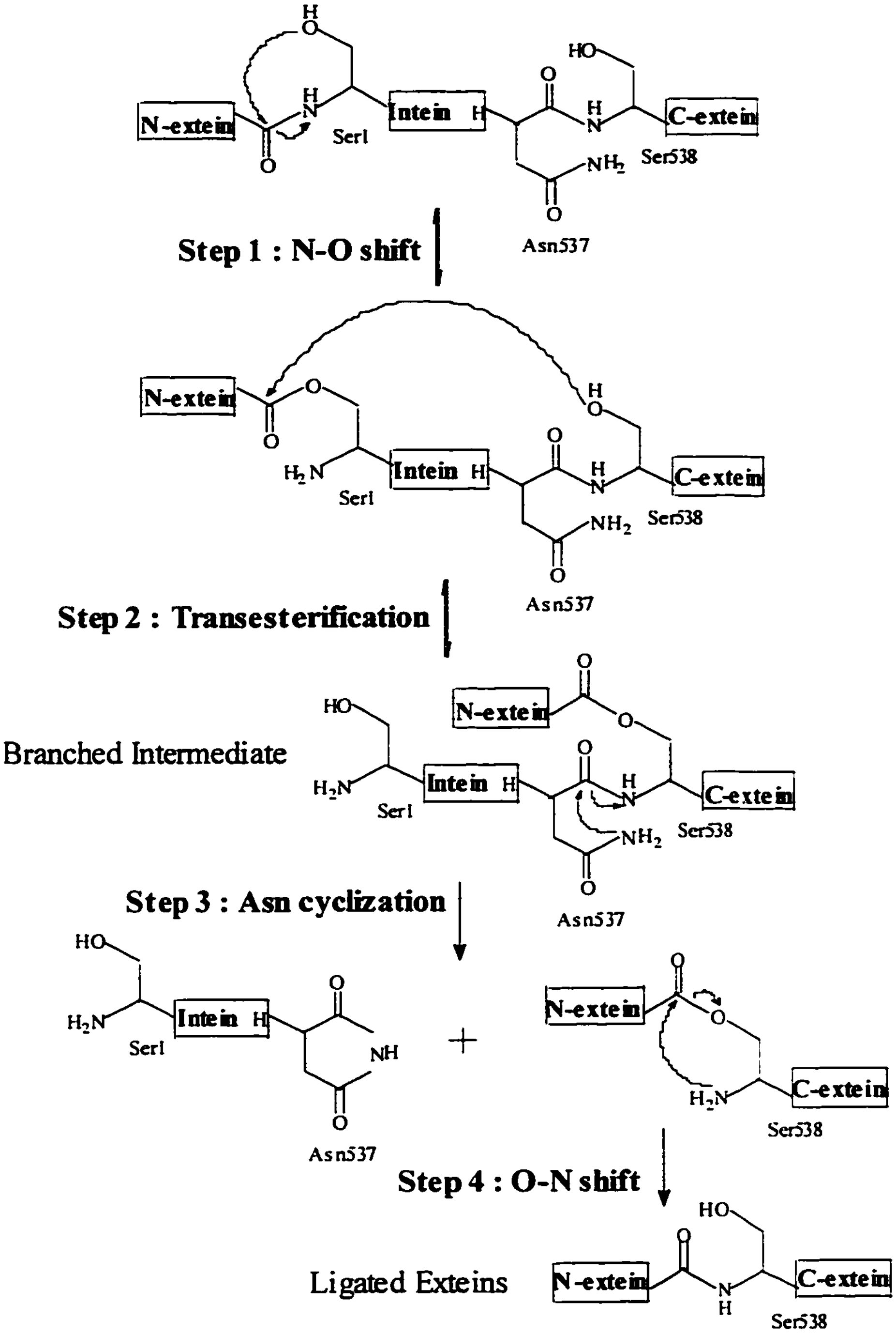

N-extein
Intein
C-extein
Serl
Ser538
Asn537
Step 1 : N-O shift
Step 2 : Transesterification
Branched Intermediate
Step 3 : Asn cyclization
Step 4 : O-N shift
Ligated Exteins

Figure 1-3

both splicing junctions, followed by nucleophilic attack of the downstream α-amino group on the upstream ester or thioester bond, and the hydrolysis of the downstream ester or thioester bond. Cooper *et al.* (Cooper *et al.*, 1993) proposed that the Asn between the intein and the C-extein caused the peptide bond cleavage, which then starts the protein splicing pathway. The side chain amino group of Asn attacks the Asn peptide carbonyl, producing a cyclic imide at the C terminus of the intein. This reaction gives peptide bond cleavage at the C-terminal of Asn. After this initiation step, a transpeptidation reaction happens between the upstream splice junction and the amino terminus of the C-extein resulting in the release of the intein and the formation of a new peptide bond between the two exteins. Another model (Clarke, 1994) suggested that the Asn at the C-terminal splice junction makes the first nucleophilic attack on the peptide bond at the N-terminal splice junction, leading to formation of a branched intermediate. The branched structure is resolved by attacks on the branched imide (or ester) from the N-terminal nucleophilic residues. This step forms an ester bond between the side chain of the N-terminal nucleophilic residue and the carbonyl derived from the hydrolyzed peptide bond. The peptide bond at the C-terminal junction is then hydrolyzed by succinimide formation by the Asn at the C-terminal junction. In the following step, the free amino group of the C-extein performs an aminolysis reaction on the ester bond at the N-terminal junction to produce the mature protein and the free intein.

Because of the lack of structural information, all these proposed pathways were based on the possible chemical reactions that the conserved junction amino acid residues might be involved in. Mutagenesis analysis of the splicing junction residues provided most of the evidence to support these hypotheses. Mutagenesis results indicated that these residues are critical for the splicing reaction. Substitution of the Cys/Ser/Thr at each junction either completely abolished splicing or retarded splicing. The invariant Asn at the C-terminal junction was also found to be essential for splicing (Davis *et al.*, 1992; Hodges *et al.*, 1992; Cooper *et al.*, 1993; Hirata *et al.*, 1992). Because of the presence of the thiol- or hydroxyl-containing residues at both extein borders, and the presence of the His-Asn-Cys/Ser/Thr at the intein-C-extein border, some of the putative chemical pathways suggested the similarity between the mechanism of a serine/cysteine protease and protein splicing. Hodges *et al.* suggested that the His-Asn-Cys (Thr/Ser) motif at an intein-C-extein junction resembles the catalytic triad in serine/cysteine proteases (Hodges *et al.*, 1992); the His activated the Ser or Cys which proceeds to attack the relevant peptide bond at the C-terminal splice junction. However, a later mutagenesis study showed that some substitution of this histidine residue still allowed partial or near wild-type levels of splicing (Cooper *et al.*, 1993). This result indicated that the His at the C-terminal splice junction might not be absolutely required for splicing. This conjecture was supported by a later discovery of naturally occurring substitutions at this position (Bult *et al.*, 1996; Wang and Liu, 1997).

A big breakthrough in protein splicing mechanism studies was the development of an *in vitro* protein splicing system by Xu and coworkers (Xu *et al.*, 1993). In this system, a fusion gene was constructed to produce a fusion protein, MIP, which consists of the maltose-binding protein (M, the N-extein), the *Pyrococcus* sp. DNA *pol* intein (I) and paramyosin (P, the C-extein). In previous studies, protein splicing was so efficient that the precursor was hardly detected. However, the intein in MIP is from an extreme thermophilic archaea growing at temperature above 95 °C. When the MIP fusion was expressed in *E. coli* at 12-32 °C, the precursor protein spliced at a slow rate. This made it possible to purify the precursor protein by amylose column and Mono Q fast protein liquid chromatography (FPLC). The ability to purify the precursor protein provided a handle on the study of the protein splicing mechanism. The *in vitro* protein splicing of the purified precursor protein enabled people to prove that protein splicing is an auto-catalytic reaction taking place at the protein level. The relatively pure precursor also excludes the requirement for *trans*-acting factors because protein splicing was detected in a reaction mixture containing only the MIP precursor. In the *in vitro* splicing reactions, the reaction rate can be controlled by temperature and pH. Splicing proceeded efficiently at temperatures of 37-65 °C, and was most efficient at the highest temperature. The rate of splicing was increased at pH 6 versus pH 7.5, and could be almost completely inhibited at pH 9 or above. This was also consistent with the hypothesis that the N-O acyl shift is an early step in splicing, because the equilibrium of N-O acyl shift favors the amide at high pH and the ester at low pH.

Based on their experimental data and earlier models, Xu *et al.* initially proposed that protein splicing starts with a serine protease-like attack by the downstream Ser on the upstream splice junction, producing an intein (I)-MP branched intermediate. Ester hydrolysis would then yield the mature protein (MP) and the free intein (I) (Xu *et al.*, 1993). Since the *in vitro* splicing of MIP takes several hours, an intermediate was detected during the time course of the reaction. The intermediate migrated at a significantly slower rate on SDS polyacrylamide gels than did the MIP precursor. Western blot analysis demonstrated that the intermediate contained the intein and both exteins. Furthermore, amino-terminal sequencing of the intermediate revealed two N-terminal sequences corresponding to the maltose-binding protein (N-extein) and the *Pyrococcus* sp. pol intein. These results indicated that this slow-migrating protein is a branched intermediate, which is reminiscent of the branched RNA intermediates formed during RNA splicing. Their subsequent experiments (Xu *et al.*, 1994) showed that the branched intermediate was alkali labile, and treating it with 6 M guanidine hydrochloride at pH 9.0 produced IP and M instead of MP and I. This result indicated that the branched intermediate consists of the N-extein (M) attached to the intein-C-extein (IP) by an alkali-labile bond, which could be either an ester or imide linkage. Evidence that supported the N-O acyl rearrangement as the initial step in protein splicing was obtained by replacing the first serine in the *Pyrococcus* sp pol intein with cysteine (Shao *et al.*, 1996). Thioester bonds are more susceptible to attack by nitrogen nucleophiles at neutral pH than an oxygen esters. As expected, the cleavage of a precursor with Ser1 to Cys1 mutation can be enhanced by nitrogen nucleophiles, suggesting that the peptide bond at the upstream splice junction undergoes an N-O/N-S acyl rearrangement. This conclusion is also supported by mutagenesis data (Xu and Perler, 1996). Furthermore, mass spectrometry revealed the presence of a succinimide ring in a carboxyl-terminal peptide (YAHN) isolated from the excised *Pyrococcus* sp. pol intein, indicating that the invariant Asn at the C terminus of intein undergoes a succinimide rearrangement (Shao *et al.*, 1995).

The mechanism proposed by Xu *et al.* is most consistent with all the reported experimental data obtained from the *in vitro* splicing system containing the thermophilic intein. A similar strategy was used in the study of the mesophilic *Saccharomyces cerevisiae* VMA intein (Chong *et al.*, 1996). It was found that the same chemical pathway is adopted in the process of protein splicing of the *Sce* VMA intein, except that N-S acyl rearrangements are involved rather than N-O acyl shifts because there are conserved cysteines at both splice junctions.

The elucidation of the mechanism of protein splicing made it possible for the scientists at New England Biolabs, Inc. to develop a protein purification system, IMPACT™. This system uses a modified *Sce* VMA intein (Chong *et al.*, 1997; Xu, 1997). This modified intein undergoes a self-cleavage reaction at its N terminus at low temperature in the presence of thiols such as DTT, β-mercaptoethanol or Cys (Fig. 1-4). A target gene is inserted into the multiple cloning site of the pTYB vector to create a fusion between the carboxyl terminus of the target protein and the amino terminus of the intein. A 5-kDa chitin-binding domain from *Bacillus circulans* has been added to the C terminus of the intein for affinity purification of the fusion protein. The expression of the three-part

Figure 1-4 Schematic illustration of cleavage reaction in the IMPACT™ system. Intein and CBD represent *Sce* VMA intein and chitin binding domain, respectively. Adapted from the IMPACT™ manual, New England Biolabs, Inc..

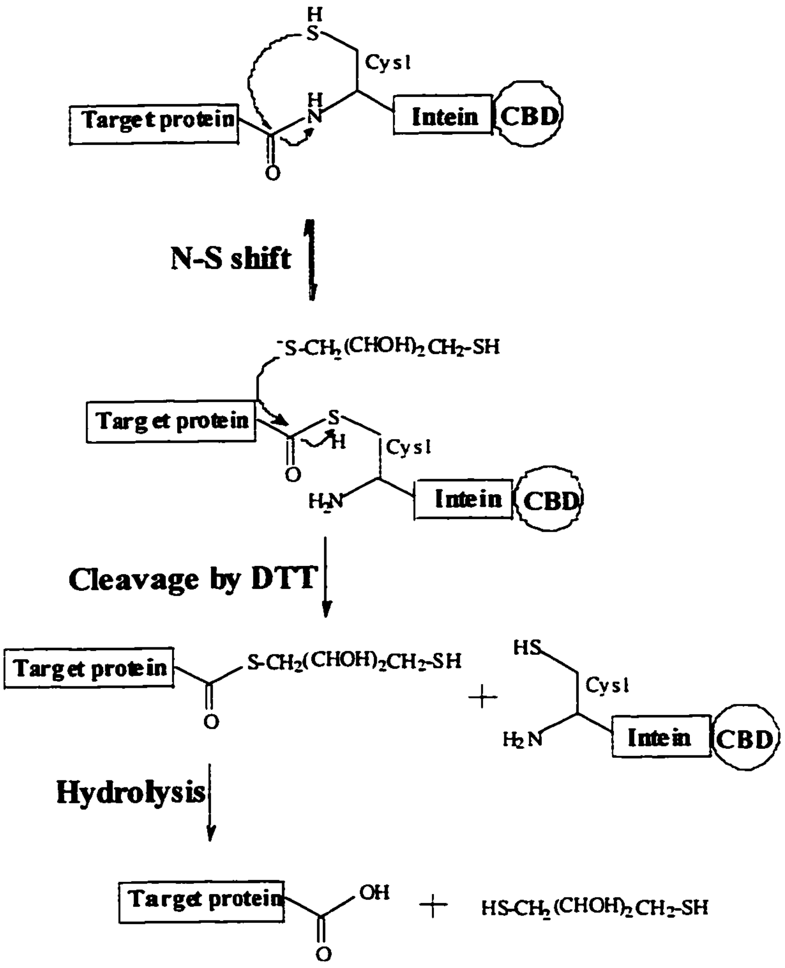

H
S
Cys1
H
N
Target protein
Intein
CBD
O
N-S shift
$^{-}$S-CH$_2$(CHOH)$_2$CH$_2$-SH
Target protein
S
H
Cys1
O
H$_2$N
Intein
CBD
Cleavage by DTT
Target protein
S-CH$_2$(CHOH)$_2$CH$_2$-SH
O
+
HS
Cys1
H$_2$N
Intein
CBD
Hydrolysis
Target protein
OH
O
+
HS-CH$_2$(CHOH)$_2$CH$_2$-SH

Figure 1-4

chimeric protein in *E. coli* is controlled by an IPTG-inducible P_{tac} or T7 promoter. When the crude cell extracts are passed through a chitin column, the fusion protein binds to the chitin column while all the contaminants are washed through the column. The fusion protein is then induced to undergo an intein-catalyzed self-cleavage on the column by overnight incubation at 4 °C in the presence of DTT. The target protein is then released into the eluant while the intein-chitin-binding domain fusion remains bound to the column.

Although the biochemical mechanism of protein splicing has been determined, it is not clear how the catalytic groups are brought together to catalyze each reaction. Information on intein structure should provide evidence needed to support the proposed models.

1.5 Structure of inteins:

The first reported crystal structure of an intein was produced with the *Saccharomyces cerevisiae* VMA intein (Duan *et al.*, 1997). The structure revealed two distinctive domains (I and II) with novel folds and corresponding to different functions (Fig. 1-5). This structure is consistent with earlier predictions based on functional studies.

Domain I of the *Sce* VMA intein is elongated and formed largely from seven β sheets. It harbors the N- and C-terminal residues and two His residues that are implicated in protein splicing. The positions of the critical junction residues are consistent with their proposed roles in the biochemical pathway of protein splicing (Hirata and Anraku, 1992; Cooper *et al.*, 1992; Chong *et al.*, 1996). Cys1 at the N-terminal splice junction and Asn454 at the C-terminal splice junction are in close proximity. This positioning is consistent with the proposal that Cys455 acts as a nucleophile to cleave the thioester at Cys1 in forming the branched intermediate. The structure analysis also revealed that two His residues, the invariant His (His79) in motif B and the conserved His at the C-terminal

Figure 1-5 Topology diagram of *Sce* VMA intein structure. The circles and triangles represent α helices and β strands, respectively. The two secondary-structure motifs related by local two-fold symmetry in domain II are enclosed in the rectangular box with dashed lines. Adapted from Duan *et al.*, 1997.

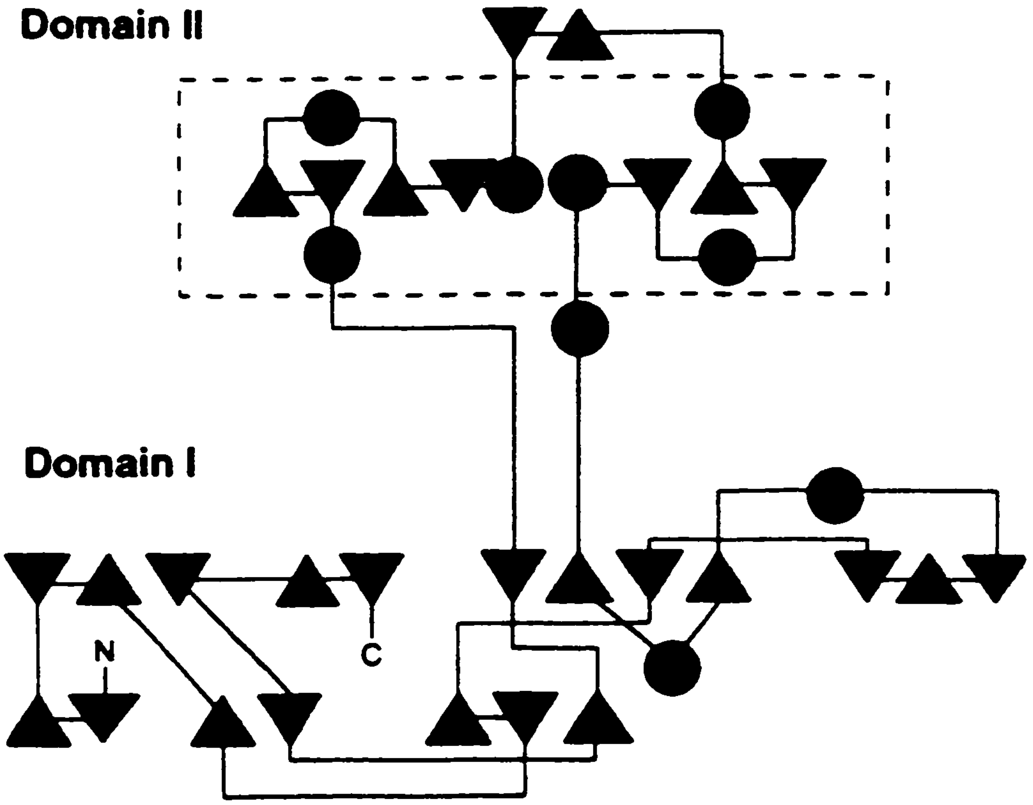


Figure 1-5

splice junction (His453), are close to the catalytic core. Due to their near-neutral pKa, the His residues may function as general acids or bases. Since His79 is close to Cys1, it may act as a proton donor/acceptor to facilitate the N/S acyl rearrangement and transesterification reactions (Pietrokovski, 1994). In fact, the imidazole side chain of His79 is situated close to Cys1. The distance between His79 and Cys455 is unknown because of the absence of Cys455 from the crystallized protein. Mutagenesis studies have suggested that His453 assists the cyclyzation of Asn-54 (Cooper *et al.*, 1993; Chong *et al.*, 1996). The structure analysis showed that the imidazole side chain of His453 is situated very close to Asn454.

In the *Sce* VMA intein, domain II is compact and is primarily composed of two similar α/β motifs related by local two-fold symmetry. It contains the putative nuclease active site with a cluster of two acidic residues and one basic residue commonly found in restriction endonucleases. The LAGLIDADG motifs form two α-helices, at the C-terminal ends of which two putative active-site Asp residues are found. The two motifs also form part of the interface between the two structurally similar subdomains. Similar structure was reported in the first crystal structure of a homing endonuclease, I-*Cre*I (Heath *et al.*, 1997). The conservation of these motifs is required for forming and maintaining the appropriate endonuclease active site geometry.

Since the structure of the *Sce* VMA intein was determined in the absence of any extein sequence, it is the structure of the excised intein. Whether the structure of the excised intein could also represent the structure of precursor protein is still unknown, because conformational changes could occur during the splicing process. Recently, the crystal structure of a 198-aa intein, the GyrA intein from *Mycobacterium xenopi*, has been resolved (Klabunde *et al.*, 1998). In this study, the intein was crystallized with one residue (Ala) of the N-extein and a Cys1-to-Ser1 substitution in the intein to capture the presplicing state. The *Mxe* GyrA intein has a compact β-structure. The conserved junction residues are situated closely on adjacent anti-parallel β stands in the central cleft of the horseshoe shaped β core, forming the catalytic center for splicing. The hydroxyl group of Ser1 is oriented toward the peptide bond and ready to attack. Thr72, Asn74 and His75 (in motif B) would assist the formation of a thioester in the native protein. When the first residue of the C-extein (Thr) is modeled in the structure, its hydroxyl group is in position to initiate the transesterification reaction. His197 can donate a proton to help the cyclyzation of Asn. However, there is no residue in this precursor positioned to deprotonate the Asn side chain. Since the endonuclease motifs are missing in the *Mxe* GyrA intein, the two splice junctions in *Mxe* GyrA intein are connected by a disordered loop. In the *Sce* VMA intein, they are connected by the endonuclease domain containing both a DNA-docking site and the catalytic center for DNA cleavage.

1.6 Inteins are mobile genetic elements:

As mentioned above, most of the inteins contain two dodecapeptide motifs. These motifs are found in the *S. cerevisiae* site-specific homing endonuclease (HO) involved in yeast mating-type switching. To initiate the mating-type switch, the HO enzyme generates a double-stranded break at a specific site in the *MAT* locus on chromosome III. In the repairing process, information at the *MAT* locus is substituted by information copied from one of two homologous loci (*HML* and *HMR*) situated near the telomeres of the same chromosome.

The dodecapeptide motifs are also shared by endonucleases encoded by open reading frames within group I self-splicing introns. Group I introns are unique in terms of their mechanism of splicing and intron mobility. The splicing of a 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) on the phosphorus atom of the 5' splicing site, forming a 3', 5'-phosphodiester bond with the first nucleotide of the intron. The phosphorus atom is then attacked by the 5' exon, resulting in ligation of the exons and the excision of the intron (Belfort, 1990; Cech, 1990). The mobility of a group I intron is site-specific. It is restricted to exchanges between alleles of genes that contain or lack the intron (Dujon, 1989). The mobility is mediated by endonucleases encoded by group I introns (Perlman and Butow, 1989; Belfort, 1990). The endonucleases recognize a site in DNA near the site of intron insertion. When an intron is present, it interrupts the site, and the DNA will be resistant to cleavage. In the absence of the intron, the endonuclease can make a double-strand break on the intron-less allele. Hence, when an empty allele (recipient locus) and an intron-containing allele (donor locus) are resident in the same cell, the endonuclease produced from the donor makes a double-stranded break in the recipient DNA. This break is then repaired by using the information on the intron-containing allele, resulting in unidirectional gene conversion with movement of the intron into the site that lacks it. This process is called intron-homing. It is responsible for maintenance of group I introns at their present positions and may also be responsible for their spread to new locations. This type of endonuclease is named homing endonucleases.

The homing endonucleases are different from type II restriction endonucleases. Homing endonucleases usually have recognition sequences that span 12-40 bp in length and contain no obvious dyad symmetry. These endonucleases can also tolerate single base-pair changes in their lengthy DNA interaction site. Many of the homing endonucleases possess the LAGLIDADG motifs. The two conserved 12-residue peptide sequences are spaced approximately 100 amino acids apart. They are not responsible for mediating substrate specificity, because each endonuclease recognizes and cleaves a different site. Instead, the residues within these motifs comprise part of the catalytic center.

One of the major suprises associated with the discovery of protein splicing is that inteins are mobile genetic elements (Gimble and Thorner, 1992). The 50-kDa intein excised from the 119-kDa primary translation product of the yeast *vma1* gene, now named as VDE (VMA-derived endonuclease), exhibits 31% amino acid sequence identity to the yeast HO endonuclease. Gimble and Thorner demonstrated that the 50-kDa intein is a site-specific endonuclease that specifically cuts at the site of insertion of the intein in a copy of the gene lacking the intein-coding sequence (Gimble and Thorner, 1992). The cleavage at the intein-less allele led to the insertion of the intein-coding sequence into this site in a heterozygous strain carrying an intein-less VMA allele and an intein-containing allele. The mechanism of this "intein homing" process is believed to be the same as that of intron homing of group I introns.

Up to now, four inteins have been demonstrated to have site-specific endonuclease activity: the *S. cerevisiae* VMA intein, the *T. litoralis* pol-1 intein, the pol-2 intein, and the *Pyrococcus* sp. pol intein. The homing ability ensures the maintenance of inteins, while their self-splicing activity assures the survival of the host, even if the intein-coding sequence is inserted into an essential gene. Inteins are likely to be selfish elements that may survive by invading and multiplying without significantly compromising their host, although it was once hypothesized that inteins could play a role in intracellular survival or pathogenesis of *Mycobacterium tuberculosis* and *Mycobacterium leprae* (Davis *et al.*, 1994).

Among all the known intein-encoded endonucleases, the *Sce* VMA intein is the one most studied. It has been shown that the endonuclease can initiate an intein homing process that transfers its DNA coding sequence to a recipient locus that lacks the intein coding sequence (Gimble and Thorner, 1992). This enzyme recognizes a 31-bp asymmetrical sequence and cuts DNA to yield 5'-phosphate and 3'-hydroxyl ends (Gimble and Thorner, 1993; Gimble and Wang, 1996). The 31-bp recognition sequence can be divided into two regions (Gimble and Wang, 1996; Wende *et al.*, 1996). Region I contains the cleavage site that is cut by the enzyme to generate a 4-bp overhang, and region II contains the adjacent 17-bp minimal binding sequence that is sufficient for high-affinity binding of enzyme. Like the other homing endonucleases, this enzyme requires Mg^{2+} as a cofactor. The metal ion is required for catalysis but not for specific binding. It can be substituted by Mn^{2+}, and the substitution leads to more efficient cleavage by the enzyme at cognate and noncognate sites (Gimble and Thorner, 1992; Wende *et al.*, 1996). The crystallography of the *Sce* VMA intein revealed the structure of the catalytic triad formed by a Lys and two Asp residues present in the endonuclease domain (Duan *et al.*, 1997). The structure is similar to the charged clusters found in restriction enzymes (Gimble and Stephens, 1995). A model describing the interaction of the endonuclease with its DNA substrate has been proposed. It suggests that the endonuclease domain contacts the cleavage site of the substrate, while the protein-splicing domain as well as the endonuclease domain are both involved in binding of the substrate (He *et al.*, 1998).

1.7 Origin and evolution of inteins:

Inteins are functionally analogous to RNA introns. Both RNA splicing and protein splicing remove intervening sequence in a gene to produce protein products different from that predicted from the DNA sequences. In contrast to RNA splicing, however, the excision of the intervening sequence in protein splicing happens at protein level rather than at RNA level. Many self-splicing introns contain protein-coding sequences. The proteins encoded by introns include the homing endonucleases. Many of the self-splicing inteins are also endonucleases similar to those encoded by group I introns. These endonucleases are believed to mediate the mobility of introns (inteins). It has been suggested that the mobile group I introns have arisen by insertion of site-specific endonuclease genes into self-splicing introns, converting them into mobile genetic elements (Belfort, 1989; Lambowitz, 1989; Loizos *et al.*, 1994). Similar events could also account for the origin of mobile inteins. They could be the result of in-frame insertion of endonuclease genes into the coding sequences of self-splicing proteins. The crystal structure of the *Sce* VMA intein provides structural information suggesting that the two functions (protein splicing and endonuclease) of inteins have evolved independently (Duan *et al.*, 1997; Klabunde *et al.*, 1998). The recent discovery of a second family of homing endonuclease other than the LAGLIDADG homing endonuclease, an HNH homing endonuclease, in the GyrB intein of *Synechocystis* also suggests the independent origins of the two functions of inteins (Dalgaard *et al.*, 1997; Pietrokovski, 1998).

The most conserved sequence motifs in inteins are those at the two splice junctions. Some of these motifs have been found in some viral proteins and in the hedgehog proteins (Pietrokovski, 1994). Hedgehog protein is a member of a family of proteins that can induce specific patterns of differentiation in a variety of tissues and structures during vertebrate and invertebrate development (Hammerschmidt *et al.*, 1997). They are synthesized as ~45-kDa precursors, which undergo autoprocessing to produce a secreted 25-kDa carboxyl-terminal fragment (Hh-C) and a 20-kDa amino-terminal fragment (Hh-N), with a cholesterol moiety covalently attached to its carboxyl-terminus (Lee *et al.*, 1994; Bumcrot *et al.*, 1995; Porter *et al.*, 1995, 1996a, 1996b). The cholesterol modification causes association of Hh-N with the cell membrane and is essential for proper Hh function. Hh-N contains all of the signaling activity of Hh proteins, while Hh-C is responsible for both the peptide bond cleavage and cholesterol transfer components of the autoprocessing reaction (Porter *et al.*, 1996a, 1996b).

Sequence analysis shows that inteins and hedgehog proteins exhibit extensive sequence similarity. A 36-aa N-extein-intein splice junction motif has been identified in the hedgehog protein from a number of organisms (Koonin, 1995). The *Drosophila* hedgehog protein has been found to undergo autoproteolysis. The cleavage site of hedgehog protein, like the upstream splicing junction in inteins, is immediately before a cysteine (Lee *et al.*, 1994). Hedgehog protein autoprocessing proceeds through two steps (Fig. 1-6). Similar to self-splicing inteins, the first step of hedgehog autoprocessing involves an intramolecular nucleophilic attack by the thiol of an absolutely conserved Cys residue on the carbonyl group of the preceding amino acid residue, forming a thioester linkage in place of the peptide bond. The second step is different from the splicing reaction of inteins where the thioester or ester bond is attacked by the hydroxyl or thiol group of a serine, threonine or cysteine residue at the C-terminal splice junction. Since no C-terminal intein splice junction motif is present in the hedgehog proteins, the nucleophile is replaced by a cholesterol molecule. In the second step, the thioester bond is subjected to a nucleophilic attack from the 3β hydroxyl group of a cholesterol molecule, resulting in a cleavage of the thioester bond, release of the Hh-C, and formation of an ester linkage between the cholesterol and the carboxyl terminus of Hh-N (Lee *et al.*, 1994; Bumcrot *et al.*, 1995; Porter *et al.*, 1995, 1996a, 1996b).

The crystal structure of a hedgehog autoprocessing domain has been determined recently (Hall *et al.*, 1997). As expected from the sequence similarity, the structure of the 17-kDa fragment of *Drosophila* Hh-C (Hh-C_{17}) is similar to the self-splicing region of inteins. Similar to what was found in the *Sce* VMA intein and the *Mxe* GyrA intein, Hh-C_{17} has an all β-strand structure with a flattened disk shape. Two structural subdomains of the Hh-C_{17} protein are related by a pseudo two-fold axis of symmetry with a single hydrophobic core.

Because of the similarity in terms of core structure, sequence, and reaction mechanism between Hh-C proteins and inteins, it has been proposed that inteins and Hh-C evolved from a common origin (Hall *et al.*, 1997). Both of them could have originated from a duplicated motif in the common ancestor. The Hh-C contains two homologous structural subdomains, which could have arisen by duplication of a primordial gene. The

Figure 1-6 Mechanism of Hedgehog protein autoprocessing. Autocatalytic processing mediated by the carboxyl-terminal domain of the Hedgehog protein precursor (Hh-C) generates an amino-terminal domain (Hh-N). The cleavage site is immediately before a Cys. Conserved amino acid sequences are found at the cleavage site, which are also present at the upstream splice junction of inteins. Reactions analogous to those involved in protein splicing may take place in the maturation of the Hedgehog protein. Adapted from Porter *et al.*, 1996b.

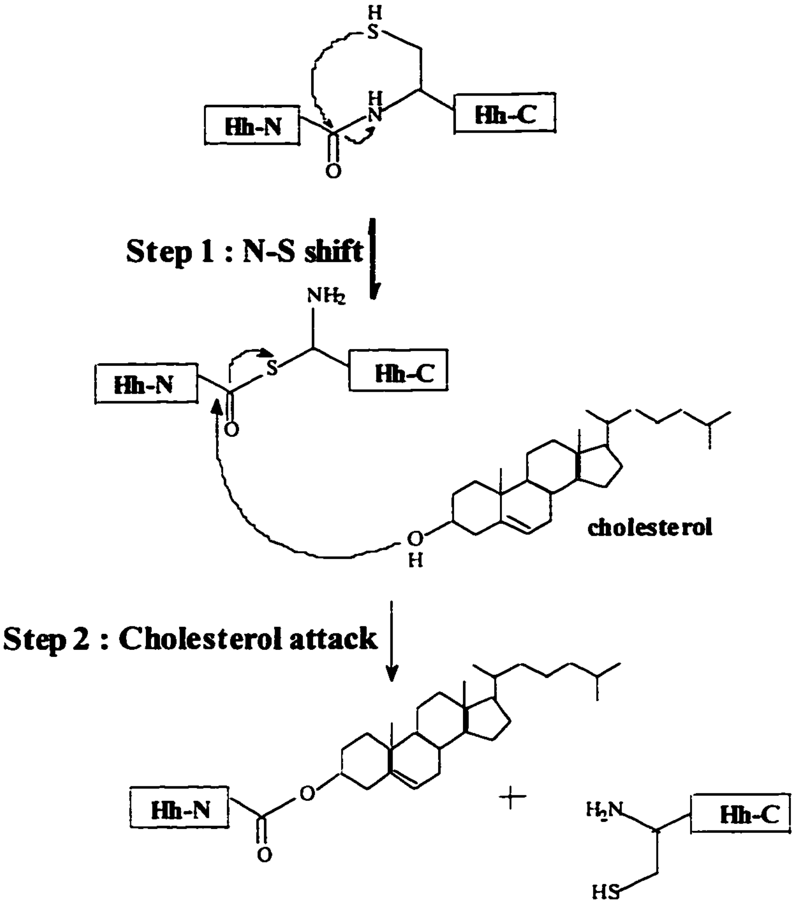

H
S
H
N
Hh-N
Hh-C
O
Step 1 : N-S shift
NH2
S
Hh-N
Hh-C
O
O
H
cholesterol
Step 2 : Cholesterol attack
Hh-N
O
O
+
H2N
Hh-C
HS

Figure 1-6

duplicated sequences do not correspond directly to the compact subdomains in Hh-C. The crystal structure of Hh-C clearly shows that the subdomains have exchanged homologous loop regions. The Hh-C module containing the internal duplication and loop swap represents the core structure conserved between hedgehog protein family members and inteins. This structure is referred to as the Hint (for Hedgehog/Intein) module (Hall *et al.*, 1997). The Hint module mediates ester/thioester formation, activating the linkage between the element and a second protein domain. Inteins and Hh-C subsequently evolved different abilities to cleave this bond. Inteins acquired the ability to ligate two exteins. Hh-C gained a sterol-recognition region directing the addition of cholesterol to the hedgehog protein signaling domain. Alternatively, the Hint module could have invaded a preexisting signaling domain or sterol-recognition region. The order of these events, including the order of module assembly, is still speculative. Inteins and hedgehog proteins could have coevolved after independent formation. Inteins also acquired a DNA recognition-region and the core endonuclease. The endonuclease activity allowed the spread of inteins by horizontal transmission. The endonuclease motifs may have been lost from some inteins. Meanwhile, Dalgaard *et al.* suggested that the hedgehog proteins originated from an intein that lost its ability to catalyze the second half of the splicing reaction but gained the ability to bind cholesterol, permitting it to react with the activated ester bond formed during the first part of the splicing reaction (Dalgaard *et al.*, 1997).

Inteins are related to the HO endonuclease in yeast *Saccharomyces cerevisiae*. HO endonuclease is a member of the dodecapeptide endonuclease family, and it is the factor responsible for mating type switch in *S. cerevisiae*. It contains all of the intein motifs except the conserved splice junction residues. It has been suggested that inteins are degenerate derivatives of HO endonuclease (Gimble and Thorner, 1992). On the other hand, HO endonuclease is most closely related to the VMA inteins found in *Saccharomyces cerevisiae* and *Candida tropicalis*. Phylogenetic analysis of these sequences, together with the inteins found in archaebacterial DNA polymerases (which are closest to the yeast VMA inteins), indicates that HO endonuclease probably arose from a homing intein at roughly the same time that *Saccharomyces* diverged from *Candida* (Keeling and Roger, 1995).

Inteins present in the same position of extein homologues from different organisms have been designated as intein alleles (Perler *et al.*, 1997). Allelic inteins are more closely related than non-allelic inteins. It is not clear whether intein alleles arose from recent intein homing events or from the acquisition of an intein by a common ancestor. If intein alleles are ancient, and some of them are lost during evolution, there must be an efficient mechanism for intein loss without inactivating the extein genes. If the intein was no longer an active endonuclease, recombination could lead to intein loss. Nevertheless, if the intein is still an active endonuclease, lateral transmission should be a more convincing hypothesis.

The genetic mobility of inteins coupled with the extremely broad phylogenetic distribution of inteins (in eukaryotes, prokaryotes, archaea) could suggest a horizontal mode of transmission (Cooper and Stevens, 1995). The RecA inteins in two different *Mycobacterium* species (*M. tuberculosis* and *M. leprea*) are not only different in size (365 amino acids versus 440 amino acids), but also have low sequence similarity: they share only 33% amino acid identity. In addition, these two inteins are found at two different positions in the RecA protein, suggesting that the two inteins arose from two independent insertion events (Davis *et al.*, 1994).

Analysis of the *gyrA* locus in different strains of various mycobacterial species revealed the presence of inteins in the GyrA proteins of four species: *Mycobacterium leprae*, *Mycobacterium flavescent*, *Mycobacterium gordonae*, and *Mycobacterium kansasii* (Fsihi *et al.*, 1996). These inteins were suggested to behave as homing endonucleases. In all four cases, the intein-coding sequences were localized at the same position in the GyrA protein. They are therefore considered to be intein alleles. The putative homing site in the *gyrA* gene appears to be occupied in some species but not in the others. Differences in the nucleotide sequences of the putative homing sites could explain the acquirsition of the intein by some strains but not by the others. This also suggested that the intein coding sequence has been acquired independently by each species since divergence from their common ancestor. In all four cases, the base composition and codon usage of the intein coding sequence is not representative of mycobacterium. The deviation of dG+dC content and codon usage between inteins and exteins also indicates that the intein coding sequences might be mobile and of foreign origin. The dG+dC content in *Mycobacterium* GyrA intein coding sequence is 47.7% while in the extein coding sequence, it is 54.2% (Fsihi *et al.*, 1996). This observation suggests horizontal transfer of the inteins.

Another piece of evidence for horizontal transmission of inteins is from the DnaB inteins of *Rhodothermus marinus* and *Synechocystis* sp. PCC6803 (Liu and Hu, 1997b). The two inteins are present at the identical position in the DnaB protein. The *Rma* DnaB intein appears to be a recent acquisition in this organism, because its codon usage and dG+dC content are quite different from the norm while the codon usage and dG+dC content of *Ssp* DnaB intein are close to the norm. The sequence identity between the two inteins are unusually high (54%), while the sequence identity between their exteins is just 37%. If two homologous inteins are not related through recent horizontal transfer, the intein sequences should have diverged much more than their extein sequences. Therefore, the *Rma* DnaB intein and the *Ssp* DnaB intein are likely to be related through recent horizontal transfer rather than through vertical inheritance. Taken together, the presence of intein alleles is more likely due to horizontal transmission rather than the early acquisition by a common ancestor.

Phylogenetic analysis of intein sequences indicates that their branching pattern reflects the evolutionary relationship between organisms rather than between extein sequences (Dalgaard *et al.*, 1997). This suggests that inteins were transported to new positions within species multiple times during evolution rather than co-evolving with the extein sequences. The insertion sites of many inteins are functionally important regions of the exteins (Dalgaard *et al.*, 1997). They appear to be inserted in parts of the extein containing residues involved in catalysis, co-factor binding or substrate binding. One hypothesis to explain this observation is that inteins inserted into such positions are under pressure to maintain their self-splicing ability so that they would not interrupt the function of host genes, while the others that inserted into less important sites would be incorporated into the exteins as an insertion.

Chapter II Materials and Methods

A. Materials

General chemicals were from BDH Inc., Anachemia, Boehringer Mannheim Biochemicals (BMB) and Sigma Chemical Company. Tris, polypeptone and yeast extract were obtained from Bethesda Research Laboratories (BRL). Lysozyme was from BMB and Phamarcia. Restriction endonucleases were from BRL, New England Biolabs, Inc. (NEB) and Promega. Klenow fragment of *E. coli* DNA polymerase I was from BRL. Nested deletion kit and dNTPs were from Pharmacia. IPTG and DTT were from ICN. SDS was from BRL. Protein marker was from NEB. T4 DNA ligase was from BRL and NEB. [α-^{32}P]dATP was from DuPont. Taq DNA polymerase was from BRL. Vent DNA polymerase was from NEB. Advantage cDNA polymerase was from Clontech. pMAL vector plasmid was from NEB, while pET-16b and pET-32 plasmids were from Novagene. Membranes used in Southern and Western blotting analysis were from Micron Separations Inc. (MSI), and PVDF membrane was from BioRad. HRP-conjugated protein marker detection pack was from NEB. HRP-conjugated S protein was from Novagene. Peroxidase-conjugated goat anti-rabbit IgG antibody was from BRL. Anti-maltose-binding-protein antibody was from NEB. Anti-thioredoxin antibody and AP-conjugated mouse anti-sheep IgG antibody were from American Diagnostica Inc. LumiGlo Chemiluminescent substrate A and B were from KPL Inc. BCIP and NBT were from BRL. TALON metal affinity resin was from Clontech. Bio-Rex 70 resin was from BioRad. Amylose resin was from NEB. IMPACT™ one step purification system was from NEB. Activated calf thymus DNA was from Sigma. Hyperfilm-ECL was from Amersham, and X-ray film was from Eastman Kodak. Oligonucleotides were from Dalton Chemical Laboratories Inc., BRL and NEB.

B. Methods

2.1 General DNA techniques:

2.1.1 Preparation of DNA:

a. Large-scale preparation of plasmid DNA (alkaline lysis method):

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 was mixed quickly by inverting, then left on ice for 5 min. Nine ml of solution III (5 M KOAc, pH 5.0) was added. The solution was mixed by inverting and left on ice for 5 min. The mixture was centrifuged at 20, 000 X g for 20 min at 4 °C. The supernatant was filtered into a new tube. DNA in the supernatant was precipitated by addition of isopropanol and centrifuged at 10, 000 X g 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).

b. 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 in an Eppendorf tube by centrifugation. The pellet was resuspended in the residual liquid remaining in the tube by vortexing. 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, and left at room temperature for 5 min. The mixture was heated in a boiling water bath for 40 sec.

The lysate was centrifuged at 14,000 rpm for 10 min to remove the fluffy pellet. DNA was precipitated by addition of 400 µl of isopropanol, incubating at room temperature for 5 min and cetrifugation at room temperature for 10 min. The DNA pellet was washed with 70% ethanol, air dried and resuspended in 50 µl TE buffer.

c. Gentle lysis method for preparation of total DNA from bacterial cells:

This method was modified from the method of Colleaux *et al.* (1986). Cells were harvested by centrifugation and rinsed in buffer A (50 mM Tris-HCl, pH 8.0, 25 % sucrose). Cells were then resuspended in 150 µl of buffer A plus 2.5 mg/ml of lysozyme, and kept on ice for 10 min. 50 µl of 0.5 M EDTA was added and the mixture was kept on ice for 5 min. 250 µl of buffer B (50 mM Tris-HCl, pH 8.0, 62 mM EDTA, 0.1% Triton X-100) was added to the mixture ice for 10 min. The mixture was extracted by phenol and chloroform/iso-amyl alcohol (24/1). DNA was precipitated by adding 1/2 volume of 7.5 M ammonium acetate and 3 volume of ethanol, chilling at -70 °C for 15 min. After spinning at 4 °C for 20 min, the DNA pellet was dissolved in 180 µl of ddH_2O. The ammonium ions were removed by another round of precipitation by adding 20 µl of 3 M sodium acetate and 500 µl of ethanol. The DNA pellet was rinsed with 70% ethanol, air dried, and dissolved in 100 µl of TE buffer.

2.1.2 DNA digestion, fragment purification:

DNAs were digested with various restriction endonucleases in the buffers recommended by enzyme suppliers. The resultant fragments were resolved by electrophoresis in a 1% agarose gel buffered by TAE buffer (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 excised. The gel pieces containing target DNA were incubated in 3 volumes of 6 M NaI at 45-55 °C for 5-10 min until the gel was completely dissolved. The solution was mixed with GlassMilk

(Bio101), chilled on ice for 5 min, then centrifuged for 5 sec at high speed to pellet the GlassMilk. The 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 incubating at 45-55 °C for 3 min.

2.1.3 DNA ligation:

DNA ligation was carried out according to the T4 DNA ligase manufacturer's instructions. Sticky-end ligations were carried out at room temperature for 1-4 hours. Blunt-end ligations were carried out at 16 °C overnight.

2.1.4 Transformation of plasmids into *E. coli* cells:

In general, *E. coli* strain DH5α is used for the cloning. *E. coli* strain BL21(DE3) is used to express fusion genes contralled by T7 promoter.

a. Preparation of competent *E. coli* cells for electroporation:

750 ml of super broth (3.2% tryptone, 2% yeast extract, 0.5% NaCl, 5 mM NaOH) was inoculated with 2 ml of *E. coli* overnight culture grown from a single colony and grown at 37 °C with vigorous shaking until A_{600} was 0.5 -1.0. The culture was chilled on ice. Cells were harvested by centrifugation at 1,000 X g at 4 °C for 5 min. Cells were washed with 750 ml, 400 ml, 100 ml and 20 ml of ice-cold 10% glycerol and finally resuspended in 2-3 ml of 10% glycerol, aliquoted and quickly frozen in a -70 °C ethanol bath. The competent cells were stored at -70 °C until use.

b. Electroporation of *E. coli* cells:

The competent cells were thawed on ice. Forty µl of cells were mixed with 1-2 µl of DNA in a cold microfuge tube, and incubated on ice for 1-5 min. The pulse generator was set to 25 µF capacitor, 1.8 kV and 200 Ω in parallel with the sample chamber. The cell-DNA mixture was transferred to a cold 0.1-cm electroporation cuvette and shaken to the bottom of the cuvette. One pulse was applied using the above settings. One ml of SOC medium (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glucose) was added to the cuvette. The cells were gently but quickly resuspended and incubated at 37 °C with shaking for 1 hour. An aliquot of this culture was spread on an LB plate containing 50 µg/ml of ampicillin and incubated at 37 °C overnight.

c. Transformation of plasmid DNA into subcloning efficiency cells:

The competent cells were purchased from BRL. The cells were thawed on ice. 50 µl of cells were mixed with 1 - 3 µl (1 - 10 ng) of DNA and incubated on ice for 30 min. Cells were heat-shocked at 37 °C for 20 sec and then put on ice for 2 min. Room-temperature LB or SOC medium (0.95 ml) was added. Cells were incubated at 37 °C with shaking for 1 hour and an aliquot was spread on an LB plate containing 50 mg/ml of ampicillin and incubated at 37 °C overnight.

2.2 Southern blot analysis:

In general, DNA was resolved on 1% agarose gel buffered with TBE buffer 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 *Hin*dIII fragments, BRL) were recorded. The gel was soaked in 0.2 M HCl for 10 min, in denaturing buffer (1.5 M NaCl, 0.5 M NaOH) for 45 min and then in neutralizing buffer (1M Tris-HCl, pH 7.5, 1.5 M NaCl) for 30 min. The gel was placed on a glass plate covered with two layers of 3MM filter paper with both ends in 20X SSC (3 M NaCl, 0.3 M Na-Citrate, pH 8.0). The gel was covered with a piece of 0.45-micron Nylon membrane and two pieces of filter paper with the same size as the gel (the membrane and the filter paper were soaked in 20X SSC before use). Dry paper towel (6-10 inches thick) was applied on top of the filter paper and left overnight. Next morning, the membrane was removed and air dried. DNA was fixed to the membrane by UV cross-linking.

DNA blots were incubated in hybridization buffer (5 X SSC, 5 X Denhardt's solution [50 X Denhardt's solution consists of 1% Ficoll, 1% polyvinylpyrolidine and 1% bovine serum albumin], 1% SDS, 100 µg/ml denatured salmon sperm DNA) at 65 °C for 4 hours before radiolabeled probes were added, and the blots were then incubated at 65 °C overnight. Blots were washed in washing buffer I (2 X SSC, 1% SDS) once at room temperature for 15 min and twice at 65 °C for 20 min each, then washed in washing buffer II (0.2 X SSC, 1% SDS) twice at 65 °C for 20 min each. Autoradiography was performed by exposing blots to an X-ray film using intensifying screens.

2.3 Probe labeling:

DNA probes were labeled by using Prime It Random Primer Kit (Stratagene) with [α-^{32}P]dATP. Twenty-five ng of template 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 5 X 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 (3,000 Ci/mmole, 10 µCi/µl), and 1 µl T7 DNA polymerase (2 U/µl) were added. The reaction mixture was incubated at 37 °C for 5 min, then 2 µl of 0.5 M EDTA was added to stop the reaction. The probes were purified by using the nucleotide removal kit (Qiagen), then made ready for hybridization by mixing the requisite amount of probe with 200 µl of hybridization buffer, heated to 100 °C for 5 min and cooled on ice.

2.4 Nested deletion:

Circular pTS1 plasmid DNA (5-10 μg) was cut with restriction endonuclease *Spe*I to generate 5'-protruding ends. This *Spe*I site located near the middle of the *Ssp* DnaB intein coding sequence. The digestion was performed in a volume to make the final DNA concentration 0.1 μg/μl. The sample was heated to 70 °C for 10 min to inactivate the enzyme. Twenty μl of digested DNA was mixed with 20 μl of 2 X exonuclease III buffer (133 mM Tris-HCl, pH 8.0, 1.33 mM $MgCl_2$, 150 mM NaCl) and equilibrated at 37 °C for 2 - 3 min. Two μl of time = 0 control was removed and mixed with 3 μl of S1 nuclease/buffer mix (33 μl of S1 buffer [150 mM potassium acetate, pH 4.6, 1.25 M NaCl, 5 mM $ZnSO_4$, 25% glycerol], 66 μl of distilled water, 40 - 60 units of S1 nuclease), set on ice. Exonuclease III (90-130 units) was added to the mixture of DNA and exonuclease III buffer, and incubated at 37 °C. Two μl of samples were removed at 1-min intervals and mixed with 3 μl of S1 nuclease/buffer mix. All the samples were set on ice until all the timed samples had been removed from the exonuclease III reaction and then incubated at room temperature for 30 min. One μl of S1 stop solution (303 mM Tris base, 50 mM EDTA) was added to the reaction mixture and heated to 65 °C for 10 min. Progressive deletion was examined by taking 3 μl of sample from each time point and electrophoresing on a 1% agarose gel. To each time-point sample, 17 μl of ligation mix (40 μl of 10 X ligation buffer [500 mM Tris-HCl, pH 7.6, 100 mM $MgCl_2$, 10 mM ATP], 80 μl of 25% glycerol, 10 units of T4 DNA ligase, 218 μl of distilled water) were added to the remaining 3 μl of sample and incubated at room temperature for 2 hours. After electrophoresis, only the samples with deletions of up to 1 kbp were used for transformation. Ten μl of sample was used to transform *E. coli* cells.

Deletion plasmids pTS1-1, pTS1-2, pTS1-4 and pTS1-5 were derived from pTS1 by the nested deletion method. The deletion junctions were confirmed by DNA sequencing.

2.5 Polymerase Chain Reaction (PCR) techniques:

2.5.1 PCR amplification of *Ssp dnaB* gene and *Ssp dnaE* gene:

To generate suitable restriction sites for cloning of *Ssp dnaB* and *Ssp dnaE* genes into expression vectors, primers were designed to carry restriction sites 4 - 6 nucleotides from the 5' ends of primers, with the remaining portions of primers forming perfect matches with the genes. PCR reactions were carried out by using the thermostable DNA polymerase Pfu (Stratagene) according to manufacturer's instructions. Generally, a 100-μl reaction mix contained 5 μl of each primer (100 ng/μl), 10 ng *Ssp* total DNA, 10 μl of 10 X *Pfu* DNA polymerase reaction buffer (200 mM Tris-HCl, pH 8.8, 20 mM $MgSO_4$, 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 1% Triton X-100, 1000 μg/ml BSA), 12.5 μl of 8 mM dNTPs, and 5 units of Pfu DNA polymerase.

To amplify the *Ssp dnaB* gene from total DNA of *Synechocystis* sp. strain PCC6803, a pair of oligonucleotide primers were used: 5'-CGGAATTCCATATGGCTGCTGCTAACCCTGCCCT and 5'-CGCTGCAGGATCCTAGTAATCATTACTTCGTTGC. PCR reaction was performed by incubating samples at 94 °C for 3 min to denature the DNA, then repeating the cycle of 55 °C for 1 min, 62 °C for 10 sec, 72 °C for 6 min, and 94 °C for 30 sec for 30 times. The last cycle was 55 °C for 1 min, 62 °C for 10 sec, and 72 °C for 10 min.

To amplify the *Ssp dnaE-n* and *dnaE-c* genes, two pairs of oligonucleotide primers were used: 5'-ATGTCCTTCGTCGGTCYTCCATATC and 5'-ATCAATAAATCGCCTTCACATTGTAATC for amplifying the 2694 bp *dnaE-n* gene; 5'-ATGGTTAAAGTTATCGGTCGTCGTTC and 5'-CTAGCCAACACTCTGGCTTTGG for amplifying the 1377-bp *dnaE-c* gene. After PCR, the products were treated with Klenow and dTTP to make sticky ends that are compatible with the expression vector pET-32, which was cut by restriction endonucleases *Eco*RI and *Hin*dIII, and then treated with Klenow and dATP.

2.5.2 PCR-based deletion method:

Deletion plasmid pTS1-3 was derived from pTS1 by a PCR-based deletion method. A pair of oligonucleotide primers was used to amplify linear DNA from the circular pTS1 plasmid in PCR reaction: 5'-GGATCCCAATTGTCACCAGAAAT-AGAAAAG and 5'-ACTCCCCAATTGTAAAGAGGAGCTTTC. A 100-μl reaction mixture contained 10 μl of each primer (10 pmol/μl), 1 ng of template plasmid DNA, 10 μl of 10 X Vent DNA polymerase buffer, 3 μl of 100 mM MgSO4, 5 μl of 4 mM dNTPs, 5 units of Taq DNA polymerase and 0.05 units of Vent DNA polymerase. The reaction was performed by repeating the cycle of 94 °C for 30 sec, 55 °C for 30 sec, and 72 °C for 10 min, for 30 times. The amplified linear DNA was digested by restriction endonuclease *Mfe*I and circularized to form pTS1-3.

2.5.3 Amplification of whole plasmids by PCR:

Generally, the template plasmid DNAs were initially treated with T4 DNA ligase to repair any nicks on the plasmids. The Advantage cDNA polymerase mix was used in the PCR reaction. The reaction was performed by incubating samples at 94 °C for 3 min to denature DNA, then repeating cycle of 55 °C for 1 min, 62 °C for 10 sec, 72 °C for 10 min, and 94 °C for 30 sec, for 5 times; cycle of 62 °C for 1 min, 72 °C for 10 min, and 94 °C for 30 sec, for 25 times. The last cycle was 62 °C for 1 min, and 72 °C for 15 min.

2.5.4 PCR-mediated random mutagenesis:

Random mutations were introduced into the *Ssp* DnaB mini-intein by PCR according to the method described previously (Cadwell and Joyce, 1992). Compared to regular PCR, the reaction mix contained a higher concentration of Mg^{2+}, and only Taq

DNA polymerase was used to increase the error rate. A 100-µl reaction mix contained 1 ng of template DNA (pMS'T2), 5 µl of each primer (100 ng/µl; 5'-ACTCCAGAATT-CCGGTCTTGTTCGATACTGTTATGG and 5'-ACAAGCCTCGAGTTAAG-AGAGAGTGGCTGCAT), 2 µl of 10 mM dGTP and dATP, 10 µl of 10 mM dCTP and dTTP, 10 µl of 10 X mutagenesis buffer (500 mM KCl, 100 mM Tris-HCl, pH 8.3, 0.1% gelatin), 14 µl 50 mM $MgCl_2$ and 5 units of Taq DNA polymerase.

2.6 Cassette replacement:

pMS'T2 contains an *Nhe*I site and an *Xho*I site flanking the N-terminal splice junction and an *Nru*I site and an *Age*I site flanking the C-terminal splice junction. These unique sites allow a convenient way of cassette substitution. pMS'T2 was digested with *Nhe*I and *Xho*I, then ligated with the complementary oligos 5'-TCGAGTGCATCAG-TGGAGATAGTTTGATCAGCTTGG and 5'-CTAGCCAAGCTGATCAAACT-ATCTCCACTGATGCAC, to create pMS'T6 (containing no native N-extein sequences). pMS'T4 was digested with *Nru*I and *Age*I, then ligated with the complementary oligos 5'-CGAATGACATCATTGTCCATAACAGTA and 5'-CCGGTACTGTTATGGACAAT-GATGTCATTCG, resulting in pMS'T8 (containing no native C-extein sequences except for the first Ser of C-extein).

2.7 SDS-polyacrylamide gels for proteins:

2.7.1 Preparative gel:

a. Separating gel:

10 ml water, 6 ml 60% sucrose, 8 ml 5 X 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.

b. Stacking gel:

5.3 ml water, 2.5 ml 4 X 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 ammonium persulfate, 5 µl TEMED.

2.7.2 Analytical gel:

a. Separating 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, 5 µl TEMED.

b. 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, 10 µl TEMED.

c. Upper tank buffer (5X):

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

d. Lower tank buffer (4X):

1.5 mM Tris-HCl, pH 8.8

2.8 Protein expression:

To express a recombinant protein in *E. coli* cells, a single colony was grown overnight at 37 °C and diluted 100 fold with fresh LB medium containing 50 µg/ml ampicillin. The culture was grown at 37 °C for 3 hours till A_{600} reached 0.5. IPTG was added at this point to a final concentration of 0.8 mM and cells were continuously grown at 37 °C for another 2-3 hours, or grown at 15 °C overnight. Cells from 1 ml culture were harvested by centrifugation at 15,000 rpm for 2 min. The cell pellet was resuspended in

100 μl of protein loading buffer (2% SDS, 62.5 mM Tris-HCl, pH 6.8. 10% glycerol, 5% 2-β mercaptoethanol, 0.05% bromophenol blue).

2.9 Protein purification:

2.9.1 Protein purification by affinity column chromatography:

a. Purification of recombinant proteins expressed from the pMAL vector:

Recombinant proteins expressed from the pMAL vector were purified by amylose column chromatography. Proteins were induced as described in section 2.8, and the induction result was verified by analytical SDS-PAGE. Cells from 100 ml of culture were harvested by centrifugation and resuspended in 10 ml column buffer (20 mM Tris-HCl. pH 7.4, 200 mM NaCl, 1 mM EDTA). The cell suspension was sonicated for 2 min. The crude extract was obtained by centrifuging at 9,000 X g for 30 min. The supernatant was loaded onto an amylose affinity column. The column was washed with 10 X volume of column buffer. The fusion protein was eluted with column buffer containing 10 mM maltose. Protein concentration was estimated by the Bradford method (Bradford, 1976).

b. Purification of recombinant proteins expressed from the pTYB (IMPACT™) vector:

The recombinant protein expressed from the TYB vector was purified by chitin column chromatography. The protein was induced as described in section 2.8 at 15 °C overnight. Cells from 1 liter of culture were harvested by centrifugation at 5,000 X g for 20 min at 4 °C. The pellet was resuspended in 50 ml of ice-cold column buffer (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 0.1% Triton X-100, 1 mM EDTA) and sonicated for 10 min on ice. The clarified extract was prepared by centrifugation at 12,000 X g for 30 min at 4 °C. Both the soluble and insoluble parts were checked on an analytical SDS-polyacrylamide gel to monitor the presence of target protein in the clarified lysate. The lysate was slowly loaded onto a chitin column. The column was washed with 10 column volumes of column buffer and quickly flushed with 3 column volumes of cleavage buffer (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 1 mM EDTA, 30 mM DTT). The flow in the column was stopped, and the column was incubated at 4 °C overnight. The target protein was eluted with column buffer the next day.

c. Purification of poly-His proteins expressed from pET vectors:

Poly-His proteins expressed from pET vectors were purified by TALON metal affinity resin. The poly-His proteins were induced as described in section 2.8. Since these proteins were all insoluble, a denaturing lysis method was used. As a pilot experiment, cells from 1-ml culture were collected by centrifugation at 14,000 rpm for 2 min. Denaturing lysis buffer (50 mM NaH_2PO_4, pH 8.0, 10 mM Tris-HCl, pH 8.0, 8 M urea [or 6 M guanidinium-HCl], 100 mM NaCl; 0.5 ml) was added to dissolve the cell pellet. The lysate was cleared by centrifugation at 14,000 rpm for 5 min. The supernatant was transferred to a tube containing 50 µl of equilibrated TALON resin. The mixture was agitated for 10 min at room temperature. The resin was pelleted by centrifugation at 14,000 rpm for 1 min and washed twice with 1 ml of denaturing lysis buffer containing 8 M urea. The bound poly-His protein was eluted by adding 50 µl of 100 mM EDTA (pH 8.0) and checked on analytical SDS-polyacrylamide gel.

Large-scale purification was also performed under denaturing conditions. The resin was equilibrated in denaturing lysis buffer. The clarified sample was added to the resin, which comprised 1/10 - 1/20 of the final volume, and gently agitated at room temperature for 20 min. The resin was pelleted at 700 X g for 5 min. The pellet was washed in 10 volumes of lysis buffer three times and in 10 volumes of wash buffer (50 mM NaH_2PO_4, pH 7.0, 8 mM urea, 100 mM NaCl) once. The poly-His protein was eluted by adding 1 resin bed volume of elution buffer (50 mM NaH_2PO_4, pH 5.0, 8 M urea, 20 mM PIPES, 100 mM NaCl).

2.9.2 Protein purification by ion-exchange column:

The DnaE protein of *Synechocystis* sp. PCC6803 was purified by the method described previously (Kim and McHenry, 1996a; Barnes and Brown, 1979) with modifications. The Bio-Rex 70 ion-exchange resin was equilibrated according to the manufacturer's instruction and packed into a 30-ml column. The *Synechocystis* sp. strain PCC6803 cells were harvested by centrifugation. The cell pellet was resuspended in buffer I (20 mM Tris-acetate, pH 8.2, 0.5 mM EDTA, 10 $MgOAc_2$) and passed twice through a French pressure cell at 15,000 psi. The suspension was clarified by centrifugation at 3,000 X g for 10 min and then at 10,500 X g for 30 min. Streptomycin was added to the supernatant at a final concentration of 0.09 g/ml. After stirring at 4 °C for 1 hour, the sample was centrifuged at 25,000 X g for 30 min. The supernatant was dialyzed against buffer II (50 mM imidazole, pH 6.5, 1 mM EDTA, 20% glycerol, 0.5 mM EDTA, 25 mM NaCl) overnight with several changes, and loaded onto the Bio-Rex 70 column. After washing the column with 10 bed volumes of buffer II, the bound protein was eluted with a 25 mM-to-400 mM NaCl gradient in buffer II. The fractions were resolved electrophoretically on an analytical SDS-polyacrylamide gel and transblotted onto nitrocellulose membrane, then analyzed by Western blot with anti-DnaE-n and anti-DnaE-c antibodies.

2.9.3 Protein purification by electroelution:

Recombinant proteins were expressed in *E. coli* cells as described in section 2.8. Total cellular proteins were resolved by electrophoresis on a preparative polyacrylmide gel. The gel was stained with staining solution (0.1% Coomassie blue R-250, 40% methanol) and destained with 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 comparing to the protein molecular weight markers on an analytical SDS-PAGE gel.

2.9.4 Antigen preparation:

Ssp DnaB intein was purified as described in section 2.9.1 b. The protein was then loaded onto a 1-ml Mono S FPLC column for further purification. The bound protein was eluted with 0-1 M NaCl gradient in 20 mM HEPES, pH 7.0.

Ssp DnaE-n and DnaE-c proteins were purified as described in section 2.9.1 c. The two proteins were resolved on a SDS-polyacrylamide preparative gel. After electrophoresis, the gel was stained in staining solution (0.1% Coomassie Blue R-250 in methanol : acetic acid : water [40/10/50]) and destained in destain solution (methanol : acetic acid : water [50/7.5/42.5]). Gel slices containing the proteins of interest were cut out and used in antibody production.

2.10 Transblotting of proteins from SDS-polyacrylamide gels to NitroPlus or PVDF membranes:

2.10.1 Transfer of proteins onto NitroPlus membranes for Western blot:

Proteins, either expressed in *E. coli* or purified from *Synechocystis* sp. strain PCC6803, were resolved by analytical SDS-PAGE. After electrophoresis, the gel was covered with a piece of NitroPlus membrane of the same size, then covered on both sides with two pieces of 3 MM filter paper. Both membrane and filter paper were soaked in transfer buffer (25 mM Tris base, 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 200 mA for 1 - 1.5 hours or at 30 volts overnight. After transfer, the membrane was submerged in block buffer (5% skim milk powder, 20 mM Tris-HCl, pH 7.5, 150 mM NaCl. 0.3% Tween-20. 0.05% Triton X-100) at room temperature with shaking for at least 1 hour.

2.10.2 Transblotting of proteins to PVDF membranes for protein micro-sequencing:

The protein purified by electroelution was resolved on a preparative SDS-polyacrylamide gel. The PVDF membrane was soaked in methanol for 1 min, then in transfer buffer (25 mM Tris base, 192 mM glycine, 10% methanol) for 5 min before use. The assembly of the transfer sandwich was the same as that for transfer to NitroPlus membranes for Western blot. The transfer was performed at 400 mA for 8 hours or 250 mA overnight at 4 °C. 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 out for protein sequencing.

2.10.3 Protein sequencing, protease digestion and peptide analysis:

Protein sequencing, protease digestion, and peptide analysis were all carried out at the Microchemistry Facility of Harvard University. The protein sample was first digested with trypsin. The resulting peptides were resolved by a C-18 HPLC column. The possible fragments which spans the splice junction were selected based on their elution positions and the UV absorption spectra. Peptides of interest were screened by mass spectrometry, and selected peptides were subjected to protein micro-sequencing.

2.11 Western blot:

The protein samples were resolved by analytical SDS-PAGE, and transblotted onto a piece of NitroPlus membrane as described in section 2.10.1. The membrane was blocked in block buffer (5% skim milk powder, 20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.3% Tween-20, 0.05% Triton-X-100) for at least 1 hour at room temperature or overnight at 4 °C. The membrane was washed in TBSTT solution (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.3% Tween-20, 0.05% Triton-X-100) for 10 min for 3 times each. The membrane was then incubated in 10 ml of TBSTT solution containing primary antibody at room temperature with shaking for 1 hour. The anti-MBP antibody is usually used with a dilution of 1:10,000, the anti--thioredoxin antibody is used with a dilution of 1:1,000, the anti-*Ssp* DnaB intein antibody is used with a dilution of 1:5,000, and the anti-*Ssp* DnaE antibody is used with a dilution of 1:2,000. After 3 10-min washings with TBSTT solution, the membrane was incubated in 10 ml of blocking buffer containing the secondary antibody (with a dilution of 1:5,000) at room temperature with shaking for 1 hour, then washed 3 times in TBSTT solution.

If the chemiluminescent detection method was used, the membrane was laid on a piece of Saran Wrap. Equal volumes of LumiGlo Chemiluminescent substrates A and B were added on the membrane and gently agitated for 1 min. The membrane was removed from the substrates solution, and wrapped in another piece of Saran Wrap and exposed to Hyperfilm-ECL film for a few sec to 30 min for appropriate signal visualization.

If the color-detection method was used, the membrane was washed briefly in AP color development buffer (100 mM NaCl, 5 mM $MgCl_2$, 10 mM Tris-HCl, pH 9.5), then incubated in 10 ml of AP color development buffer containing 33 µl of BCIP and 44 µl of NBT at room temperature till the color development was complete. The development was stopped by washing the membrane in distilled water for 10 min with gentle agitation. The membrane was air dried.

2.12 Endonuclease activity assay:

2.12.1 Construction of homing site for *Ssp* DnaB intein:

The intein insertion site for the *Ssp* DnaB intein was constructed by using two complementary oligo nucleotides: 5'-TCGAGATGTCAGATTTAAGAGAGAGTGGCAGTATCGAACAAGACGCAGA-TTTAG and 5'-GATCCTAAATCTGCGTCTTGTTCGATACTGCCACTCTCTC-TTAAATCTGACATC. The cassette formed by these two oligos was inserted into plasmid Litmus 28 at *Xho*I and *Bam*HI sites. The resulting plasmid, pLSH2, contains a 48-bp of sequence corresponding to the insertion site of the *Ssp* DnaB intein coding sequence.

2.12.2 *In vitro* endonuclease activity assay:

Plasmid pLSH2, containing the 48-bp homing site for the *Ssp* DnaB intein, was linearized by cutting with restriction enzyme *Sca*I. The linear DNA (100 ng) was then incubated with purified *Ssp* DnaB intein (5 µg) in 50 µl buffer containing various concentrations of NaCl and Mg^{2+} (Buffer I:10 mM Tris Propane-HCl, pH 7.0, 10 mM $MgCl_2$, 1 mM DTT; Buffer II: 10 mM Tris-HCl, pH 7.9, 10 mM $MgCl_2$, 50 mM NaCl, 1 mM DTT; Buffer III: 50 mM Tris-HCl, pH 7.9, 10 mM $MgCl_2$, 100 mM NaCl, 1 mM DTT; Buffer IV: 20 mM Tris-acetate, pH 7.9, 10 mM magnesium acetate, 50 mM potassium acetate, 1 mM DTT; Buffer V: 10 mM Tris-HCl, pH 8.6, 10 mM $MgCl_2$, 150 mM NaCl, 1 mM DTT) at 37 °C for 2 hours. One µl of proteinase K (30 mg/ml) was added to the digests and incubated at 37 °C for 30 min. 25 µl of each digest were loaded onto an 1% agarose gel. After electrophoresis, the gel was stained with EtBr and the bands were visualized under UV light.

2.12.3 *In vivo* endonuclease activity assay:

The *in vivo* endonuclease assay was carried out using the method described previously (Gauthier *et al.*, 1991) with modifications. The 286-bp *Bgl*II/*Pvu*II fragment from pLSH2 containing the 48-bp homing site for the *Ssp* DnaB intein was cloned at the

*Eco*RI site in pTS1, which contains the *Ssp dnaB* gene, to form pHC-194. The protein was induced as described in section 2. 8. One ml of cells were sampled after the induction had continued for 10 min, 30 min, 1 hour, 2 hours and 3 hours. Meanwhile, a control experiment (without induction by IPTG) was performed under same conditions, and samples were removed at the same time points. The total DNA was extracted as described in section 2.1.1 c. After treating the total DNA with restriction enzyme *Pst*I, the DNA was loaded onto a 1% agarose gel. After electrophoresis, the DNA was transferred to a nylon membrane. The bands were detected by Southern blot analysis using *Pst*I-linearized pHC-194 as probe.

Chapter III Characterization of a cyanobacterial DnaB intein

3.1 Introduction:

A number of short sequence motifs (sequence motifs A to H) in inteins have shown a low but significant degree of conservation among inteins (Pietrokovski, 1994; Perler *et al.*, 1997). These short sequence motifs, combined with other criteria such as in-frame insertion in a gene, can be used to find new inteins. In this study, these criteria were used to find a new intein sequence in a blue-green alga, cyanobacterium *Synechocystis* sp. PCC6803. This intein, named *Ssp* DnaB, is found in the DnaB protein of this organism. DnaB is a DNA helicase. In *E. coli*, DnaB protein is one of the essential components of chromosome replication (Kornberg, 1980). It participates in the initiation and elongation stages of replication (Zyskind and Smith, 1977; Wechsler and Gross, 1971). DnaB protein is composed of at least two discrete domains: the N-terminal domain is essential for interactions of the protein with other replication proteins; the C-terminal domain is responsible for DNA-dependent ATPase activity and nucleotide binding (Nakayama *et al.*, 1984). In *Ssp*, a 429-aa intervening sequence is found to be inserted in a highly conserved region of DnaB protein. This 429-aa region contains all the known intein motifs. This putative *Ssp* DnaB intein was demonstrated to undergo protein splicing when tested in *E. coli* cells, indicating it is an active intein.

Although several inteins have been shown to be able to splice in a foreign context other than their native exteins, the role of extein sequences in protein splicing is still under investigation. Protein splicing mechanism studies indicated that protein splicing involves amino acid residues at the two splice junctions. These residues are mainly in sequence motifs A and G. They are separated from the two endonuclease motifs (motifs C and E) in the primary sequence by more than 100 amino acids. This spacing raises the question of whether the two functions of inteins are functionally and structurally separable. Previous studies have shown that inactivation of the endonuclease activity would not affect intein's splicing activity (Hodges *et al.*, 1992). However, it is not known whether the whole intein sequence (including the functionally unrelated endonuclease motifs) is required for the proper folding of the precursor protein for efficient splicing. In this study, I first demonstrated that the *Ssp* DnaB intein is capable of protein splicing with or without its native exteins when tested in *E. coli* cells. Subsequently, a centrally located 275-aa sequence of this intein, which corresponds to the entire endonuclease domain, was deleted without affecting the intein's splicing activity. The resulting mini-intein was further split into two fragments, and efficient protein *trans*-splicing was observed. These results indicate that the N- and C-terminal regions of the *Ssp* DnaB intein, whether physically linked or not, can come together to form a protein-splicing domain that is functionally sufficient and structurally independent from the centrally located endonuclease domain. Random mutagenesis in the intein sequence revealed the importance of some internal residues for protein splicing.

Inteins are bi-functional elements, possessing both protein splicing and endonuclease activities (Gimble and Thorner, 1992; Doolittle, 1993; Belfort and Roberts, 1997). Many inteins have sequence motifs resembling the LAGLIDADG motifs of intron-encoded endonucleases, and some of them have been demonstrated to have site-specific endonuclease activity, which is believed to mediate intein mobility by using similar mechanism as intron homing (Shub and Goodrich-Blair, 1992; Gimble and Thorner, 1993). The *Ssp* DnaB intein was purified and its endonuclease activity was tested on an artificial homing site for this intein.

3.2 Results:

3.2.1 Sequence analysis of the DnaB intein of *Synechocystis* sp strain PCC6803:

In a Blast search of GenBank for homologues of the *Porphyra purpurea* chloroplast DnaB protein, a putative intein coding sequence was noticed in the *dnaB* gene of the cyanobacterium *Synechocystis* sp. strain PCC6803. The *dnaB* gene (GenBank Accession No. D64003) was reported as part of the complete genome sequence of this organism (Kaneko *et al.*, 1996). Further analysis of this gene shows that the predicted *Ssp* DnaB protein is much larger than its *E. coli* homologue, and this is due to the presence of a large intervening sequence (429 amino acid residues long) in *Ssp* DnaB (Fig. 3-1). Considering only the extein sequences, the *Ssp* DnaB protein is 36%, 33%, and 36% identical to the DnaB proteins of *Rhodothermus marinus*, *P. purpurea* chloroplast, and *E. coli*, respectively (Fig. 3-2). These levels of sequence identity are comparable to the 35% sequence identity between the DnaB proteins of *R. marinus* and *E. coli*. Therefore, the predicted *Ssp* DnaB protein is a homologue of known DnaB proteins.

The *Ssp* DnaB protein clearly contains an intein. This intein, positioned between residue 381 and residue 809, interrupts a 14-aa stretch of sequence that is extremely conserved among DnaB proteins, so that the intein sequence boundaries are easily defined (Fig. 3-2). The predicted intein boundaries also agree with other intein-defining features, including a nucleophilic residue (Cys) at the N terminus of the intein, a His-Asn dipeptide at the C terminus of the intein, and another nucleophilic residue (Ser) at the beginning of C-extein. These four residues and their positions are highly conserved among known inteins and are known to be critical for the chemistry of protein splicing (Shao *et al.*, 1995, 1996; Xu *et al.*, 1994). In addition, the *Ssp* DnaB intein contains putative intein sequence motifs (motifs A to H) that are significantly conserved among known inteins (Fig. 3-3).

Figure 3-1 Schematic illustration and comparison of DnaB proteins of *Synechocystis* sp. PCC6803 (*Ssp*), *Rhodothermus marinus* (*Rma*), *Porphyra purpurea* chloroplast (*Ppu*), and *E. coli* (*Eco*). Hatched box and solid box represent extein and intein, respectively. Number of residues are shown for the *Ssp* DnaB extein and intein sequences.

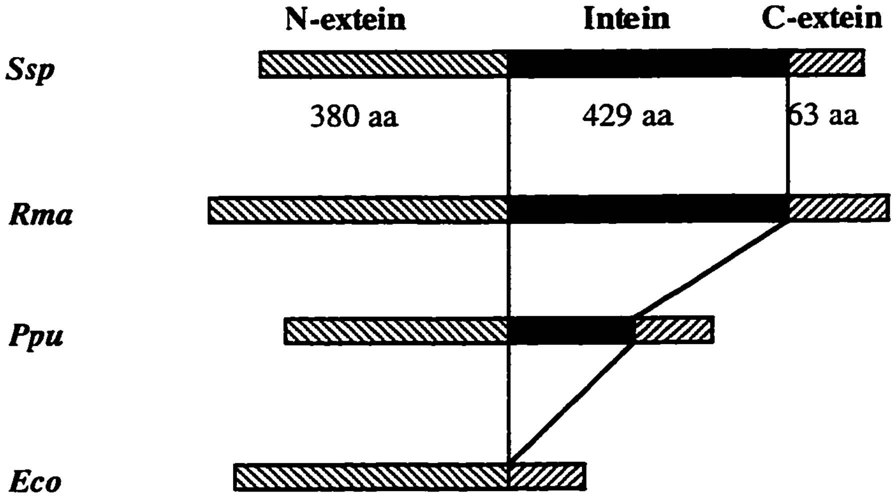


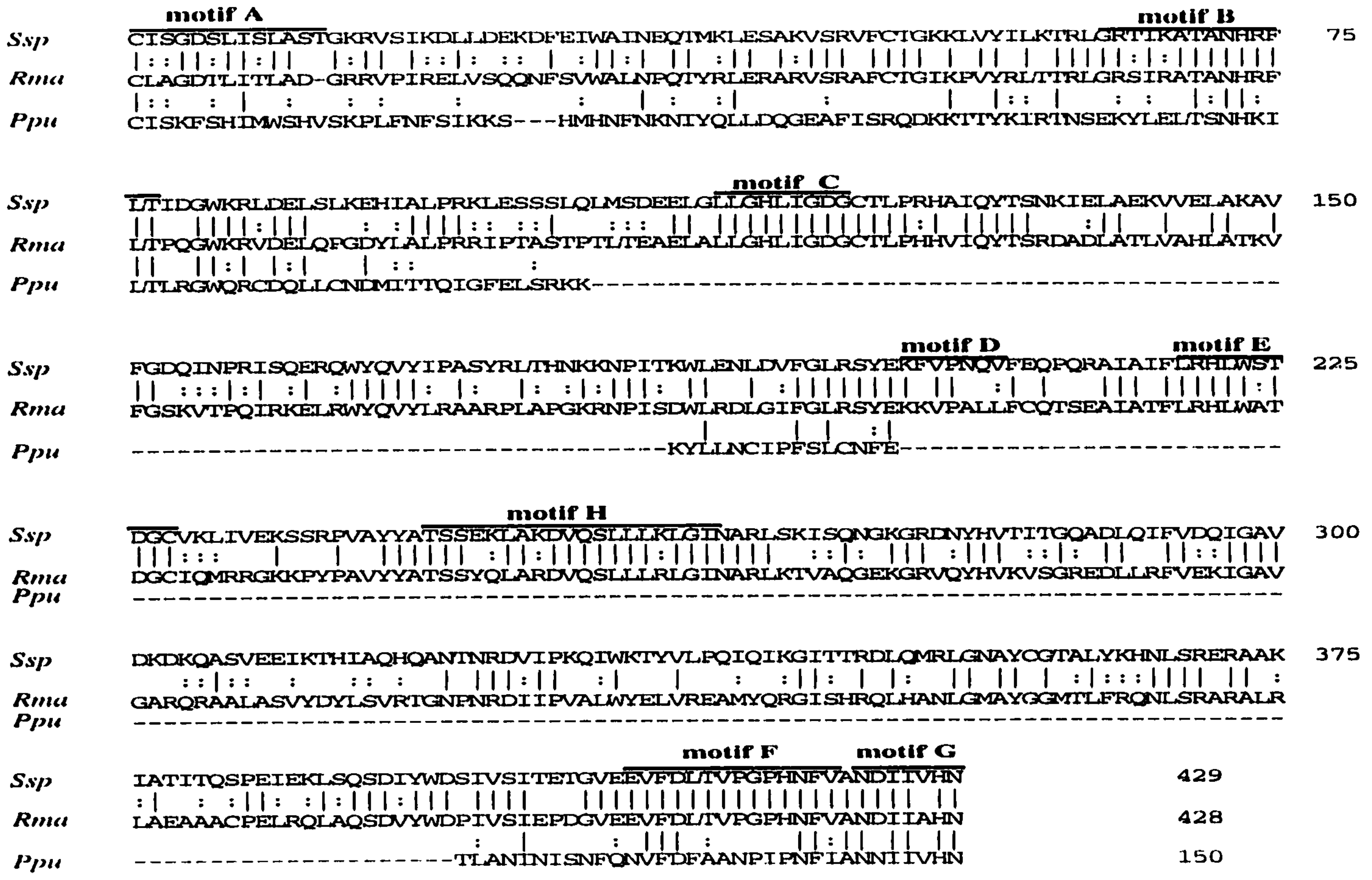
Figure 3-1

Figure 3-2 Comparison of amino acid sequence of DnaB proteins of *Synechocystis* sp. PCC6803 (*Ssp*), *Rhodothermus marinus* (*Rma*), *Porphyra purpurea* chloroplast (*Ppu*), and *E. coli* (*Eco*). The total numbers of residues in the sequences are shown at the end of the corresponding sequences. An intein is marked by a letter I in parentheses. Putative sequence motifs are marked by lines, including an ATP-binding motif (a), a DNA-binding motif (b), and a leucine-zipper motif (c). Hyphens represent gaps introduced to optimize the alignment; * and . mark positions of identical and similar amino acids, respectively.

```
Ssp     --------------------------------------MAANPALPPQNIEAEECILGGI
Rma     MAEFEERPRLSIGEEEAPPYPLEKLTGGRRTRAQIHALHQQAGRVPPQAVELEQAVLGAM
Ppu     ----------------------------------------------------------ML
Eco     -------------------MAGNKPFNKQQAEPRERDPQVAGLKVPPHSIEAEQSVLGGL
                                                                   .

Ssp     LLDPEAMGRIIDLLVVDAFYVKAHRLIYEAMLSLHGQSQPTDLMSVSSWLQDHHHFEAIG
Rma     LIEPEAIPRALEILTPEAFYDGRHQRIFRAIVRLFEQNRGVDLLTVTEELRRTGELEQAG
Ppu     TQESEDLLKQIEKLSPDFFYFKSNSLVYRAILETVNPIDKIALVSLLTALNTNNLIRQLG
Eco     MLDNERWDDVAERVVADDFYTRPHRHIFTEMARLQESGSPIDLITLAESLERQGQLDSVG
          . *      . . .  **   .  ..  .           *...   *         *

Ssp     GMVKLTQLLDRTISAVNIDRFAALIMDKYLRRQLIAAGHDIVDLGYETS-KELETIFDES
Rma     DTIYLSELTTRVASAANVEYHARIIAEKLLRR-MIEVMTLLVGRAYDPA-ADAFELLDEV
Ppu     RLETIMKLIENSPASNIIYEYSKVILDNYVKRLLLKSGDSLCLISCSKK-QITQSVITSV
Eco     GFAYLAELSKNTPSAANISAYADIVRERAVVREMISVANEIAEAGFDPQGRTSEDLLDLA
            .  *     ..  .   . ..  .  . * ..     .              .

Ssp     EQKIFRLTQSRPQ--AGLVPLSETLVNTFIELDKLHEKLSS--PGVETQFYDLDAMTGGL
Rma     EAEIFRLSDVHLR--KAARSMNEVVKETLERLEAIHGRPGG-ITGVPSGFHQLDALTGGW
Ppu     ASQLTIAYEILED--EGTYTLAEIFASLLVSLDTKKKISIN--SGIFSGFWQLDLITNGF
Eco     ESRVFKIAESRANKDEGPKNIADVLDATVARIEQLFQQPHDGVTGVNTGYDDLNKKTAGL
           .                .  .        ..           *. . .  *   * *

                  a
Ssp     QRADLIILAGRPSMGKTAFGLGIAANIAK-NQNLPVAIFSLEMSKEQLALRLVASESLID
Rma     QRGDLIIIAARPSMGKTAFALSCRNAALHPHYGTGVAIFSLEMGAEQLAQRLLTAEAASM
Ppu     QKSDLIIIAGRPSMGKTAFAINITRHIIK-TSQYYVILFSLEMSTEQLLRRILAQECHLN
Eco     QPSDLIIVAARPSMGKTTFAMNLVENAAM-LQDKPVLIFSLEMPSEQIMMRSLASLSRVD
        *  ****.* *******.* .              * .*****  **.  * ..  .
                                                   b
Ssp     SNRLRTGHFSQAEFEPLTAAMGTLSS-LPIYIDDTASISVTQMRSQVRRLQSEQKGPLGM
Rma     PR--RPAPDGCATRTGVSWPARRPLSDAPIFIDDTPSLGVLELRAKCRRLKAEHD--IGL
Ppu     SQKIQSGQLTNVEWQRIVEESKILAN-LNFYIDDSAEISCDIIKVKVKLLRLQGKK-IKL
Eco     QTKIRTGQLDDEDWARISGTMGILLEKRNIYIDDSSGLTPTEVRSRARRIAREHGG-IGL
            .           .             .***.  .    .. . . .  .    . .
                                             c
Ssp     VLIDYLQLMEG----GSDNRVQELSKITRSLKGLAREINAPVIALSQLSRAVESR-TNKR
Rma     VIVDYLQLMQASHMPRNANREQEIAQISRSLKALAKELNVPVVALSQLSRAVETRGGDKR
Ppu     IIIDYLQLLQES--KKSENRSQELSLITRSLKILARELNLPILVLSQLNRNLESR-HNKR
Eco     IMIDYLQLMRVP--ALSDNRTLEIAEISRSLKALAKELNVPVVALSQLNRSLEQR-ADKR
        ...*****.         **  *.. *.**** **.*.* *.. **** * .* *   **

Ssp     PMMSDLRESG(I)SIEQDADLIMMIYRDEYYNPDTPD-----PGVAELLIVKHRNGPTGV
Rma     PQLSDLRESG(I)SIEQDADVVLFIYRPERYGITVDENGNPTEGIAEIIIGKQRNGPTGT
Ppu     PLLSDLRESG(I)SIEQDADLVIMLYRESYYNKEMEM-----EDMTEIIVAKHRNGPLGT
Eco     PVNSDLRESG---SIEQDADLIMFIYRDEVYHENSDL-----KGIAEIIIGKQRNGPIGT
        *  *******   *******..  .**   *             ..*... *.**** *

Ssp     VKLLFKPEFTQFLNLQRSNDY                                    872
Rma     VRLAFINQYARFENLTMYQPEPGTPLPETPDETILPSGPPDEAPF            941
Ppu     FQLKFDANLANFLNV                                          686
Eco     VRLTFNGQWSRFDNYAG-PQYDDE                                 471
         .* *      * *
```

Figure 3-2

Figure 3-3 Comparison of DnaB intein sequences of *Synechocystis* sp. PCC6803 (*Ssp*), *Rhodothermus marinus* (*Rma*), *Porphyra purpurea* chloroplast (*Ppu*). The *Synechocystis* sp. PCC6803 (*Ssp*) sequence is numbered throughout, while the total numbers of residues in each sequence is shown at the end of that sequence. Putative intein motifs are marked. Hyphens represent gaps introduced to optimize the alignment; | and : mark positions of identical and similar amino acids, respectively.



```
                motif A                                                          motif B
Ssp   CISGDSLISLASTGKRVSIKDLLDEKDFEIWAINEQIMKLESAKVSRVFCTGKKLVYILKTRLGRTIKATANHRF   75
Rma   CLAGDTLITLAD-GRRVPIRELVSQQNFSVWALNPQTYRLERARVSRAFCTGIKPVYRLTTRLGRSIRATANHRF
Ppu   CISKFSHIMWSHVSKPLFNFSIKKS---HMHNFNKNIYQLLDQGEAFISRQDKKTTYKIRTNSEKYLELTSNHKI

                                                  motif C
Ssp   LTIDGWKRLDELSLKEHIALPRKLESSSLQLMSDEELGLLGHLIGDGCTLPRHAIQYTSNKIELAEKVVELAKAV   150
Rma   LTPQGWKRVDELQFGDYLALPRRIPTASTPTLTEAELALLGHLIGDGCTLPHHVIQYTSRDADLATLVAHLATKV
Ppu   LTLRGWQRCDQLLCNDMITTQIGFELSRKK---------------------------------------------

                                                            motif D         motif E
Ssp   FGDQINPRISQERQWYQVYIPASYRLTHNKKNPITKWLENLDVFGLRSYEKFVPNQVFEQPQRAIAIFLRHLWST   225
Rma   FGSKVTPQIRKELRWYQVYLRAARPLAPGKRNPISDWLRDLGIFGLRSYEKKVPALLFCQTSEAIATFLRHLWAT
Ppu   -----------------------------------KYLLNCIPFSLCNFE-------------------------

                               motif H
Ssp   DGCVKLIVEKSSRPVAYYATSSEKLAKDVQSLLLKLGINARLSKISQNGKGRDNYHVTITGQADLQIFVDQIGAV   300
Rma   DGCIQMRRGKKPYPAVYYATSSYQLARDVQSLLLRLGINARLKTVAQGEKGRVQYHVKVSGREDLLRFVEKIGAV
Ppu   ---------------------------------------------------------------------------

Ssp   DKDKQASVEEIKTHIAQHQANTNRDVIPKQIWKTYVLPQIQIKGITTRDLQMRLGNAYCGTALYKHNLSRERAAK   375
Rma   GARQRAALASVYDYLSVRTGNPNRDIIPVALWYELVREAMYQRGISHRQLHANLGMAYGGMTLFRQNLSRARALR
Ppu   ---------------------------------------------------------------------------

                                           motif F       motif G
Ssp   IATITQSPEIEKLSQSDIYWDSIVSITETGVEEVFDLTVPGPHNFVANDIIVHN                        429
Rma   LAEAAACPELRQLAQSDVYWDPIVSIEPDGVEEVFDLTVPGPHNFVANDIIAHN                        428
Ppu   ---------------------TLANINISNFQNVFDFAANPIPNFIANNIIVHN                        150
```

Figure 3-3

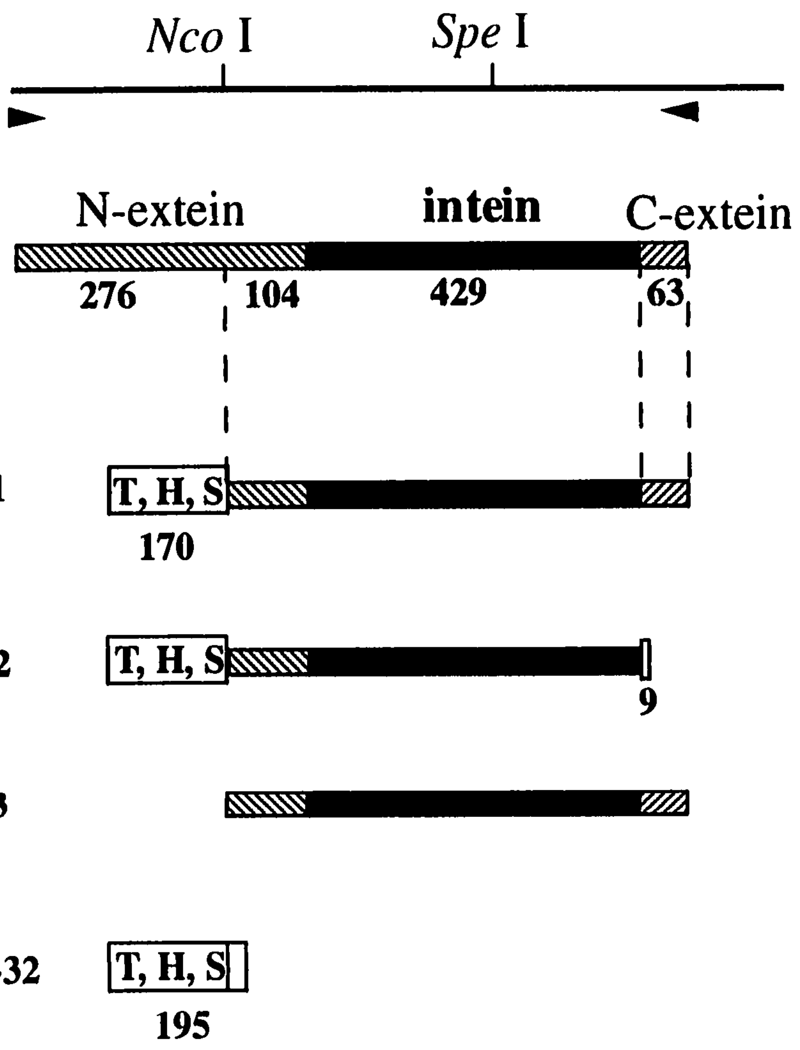
The *Ssp* DnaB intein is similar in size to the 428-aa *Rma* DnaB intein but much larger than the 150-aa DnaB intein of the *P. purpurea* chloroplast. Sequence comparison between DnaB inteins of *Synechocystis* sp. PCC6803 and *R. marinus* revealed a 54% sequence identity and a 74% sequence similarity (Fig. 3-3). The much shorter DnaB sequence of the *P. purpurea* chloroplast corresponds mostly to the two terminal regions of the *Ssp* DnaB intein, and the 150-aa sequence is 29% and 26% identical to corresponding DnaB intein sequences of *Synechocystis* sp. PCC6803 and *R. marinus*, respectively. However, all three DnaB inteins are positioned identically in their respective DnaB proteins. The three DnaB inteins are therefore considered as intein alleles.

3.2.2 Demonstration of protein splicing with the *Ssp* DnaB intein in *E. coli*:

The *Ssp* DnaB intein was tested for protein splicing activity in *E. coli* cells. The complete *Ssp dnaB* gene was amplified from total DNA of *Synechocystis* sp. strain PCC6803 by polymerase chain reaction (described in section 2.5.1). The entire *Ssp dnaB* gene could not be cloned in an expression plasmid vector, presumably due to toxicity of the gene product (a DNA helicase) to the *E. coli* cell. On the other hand, clones containing partial *Ssp dnaB* genes were readily obtained, including three recombinant plasmids (pTS1, pTS2 and pTS3) that encode fusion proteins consisting of the complete intein sequence flanked by various amount of extein sequences and tag sequences (Fig. 3-4). Recombinant plasmid pTS1 was constructed by inserting a 1796-bp *Nco*I-*Bam*HI DNA fragment (blunt ended) of the *dnaB* gene into the expression vector pET-32 at its *Bam*HI site (blunt ended), so that the partial *dnaB* gene was in-frame with the upstream vector-encoded coding sequence. Recombinant plasmid pTS2 was constructed by removing a 174-bp fragment from the 3' end of the *dnaB* gene. pTS3 was constructed by inserting the 1796-bp *Nco*I-*Bam*HI DNA fragment of the *dnaB* gene into the expression plasmid vector pET-16b at its *Nco*I-*Bam*HI site. In recombinant plasmids pTS1 and pTS2, a thioredoxin gene, a poly-His coding sequence and an S-tag coding sequence, all encoded in the vector, are added to the N terminus of the N-extein. This situation makes it possible to purify the fusion proteins by using metal affinity resin. The fusion proteins can also be identified by using the HRP conjugated S protein, which specifically recognizes the 15-aa S-tag sequence, in Western blot analysis. In all these recombinant plasmids, expression of the resultant fusion gene is controlled by an IPTG-inducible T7 promoter.

Each of the recombinant plasmids was introduced into *E. coli* cells to produce the corresponding fusion protein and to observe possible protein splicing products (Fig. 3-5). In cells containing plasmid pTS1, three protein products were observed. Their sizes corresponded well with the predicted sizes of a precursor protein (86 kDa), a spliced protein (37 kDa), and an excised intein (49 kDa). In cells containing plasmid pTS2, three protein products were also observed, and their sizes corresponded well with the predicted sizes of a precursor protein (80 kDa), a spliced protein (31 kDa), and an excised intein (49 kDa). Similarly, cells containing plasmid pTS3 produced three proteins corresponding to a precursor (68 kDa), a spliced protein (19 kDa), and an excised intein (49 kDa). In addition to identification by size, the precursor protein and spliced protein bands in cells containing plasmids pTS1 and pTS2 were further identified by their selective binding to metal affinity resin (a property of the poly-His tag) and to the S protein (a property of the S tag). The intein band was identified by its size, by the fact that its size was not affected by changing the extein sequences, and also by Western blot analysis using antibody raised against this intein. For construct pTS1, significant amount of the precursor protein was accumulated while little spliced protein was detected when the proteins were induced at 37 °C. The incompleteness of the splicing reaction in this situation may be caused by misfolding and precipitation of the precursor protein, because a major portion of the precursor protein was in the insoluble fraction, indicating formation of inclusion bodies. When the induction was carried out at lower temperature (15 °C - 25 °C), significant amount of spliced protein was observed, although some precursor protein was still detectable (Fig. 3-6). Lower induction

Figure 3-4 Schematic illustration of fusion proteins constructed for demonstration of protein splicing by the *Ssp* DnaB intein in *E. coli* cells. On the top line, arrow heads represent oligonucleotide primers used in cloning the *Ssp* DnaB coding sequence. Restriction sites used in this study are also shown. DnaB intein (solid box) and extein (hatched box) sequences are fused with vector-encoded sequences (open box), with the number of residues shown for each sequence domain. On the vector(pET-32)-encoded sequence, T, H, and S stand for thioredoxin protein, poly-histidine tag, and S tag, respectively. For each construct (pTS1, pTS2, pTS3, and pET-32 as control), calculated molecular weights are listed for the predicted precursor protein, the excised intein and the spliced protein.



	Precursor protein	Excised intein	Spliced protein
pTS1	86 kDa	49 kDa	37 kDa
pTS2	80 kDa	49 kDa	31 kDa
pTS3	68 kDa	49 kDa	19 kDa
pET-32	21 kDa		

Figure 3-4

Figure 3-5 Production of the fusion proteins and protein splicing of *Ssp* DnaB intein in *E. coli* cells. *E. coli cells* containing the specified plasmid were induced at 25 °C by adding IPTG. The induced proteins were analyzed by SDS-polyacrylamide gel electrophoresis and visualized by Coomassie Blue staining. Proteins in lanes 1-4, 5-8, 9-11, 12-13 are from plasmids pET-32, pTS1, pTS2, and pTS3, respectively. Lanes 1, 5, 9, and 12: before induction. Lanes 2, 6, 10, and 13: after induction. Lanes 3, 7, and 11: proteins isolated by using metal affinity resin that recognizes the poly-histidine tag. Lane 4: Western blot analysis using the S protein (HRP-conjugated) that recognizes the S tag. Lane 8: Western blot analysis using the antiserum raised against *Ssp* DnaB intein. Letters P, I, S, and Trx mark positions of precursor protein, excised intein, spliced protein, and thioredoxin, respectively.

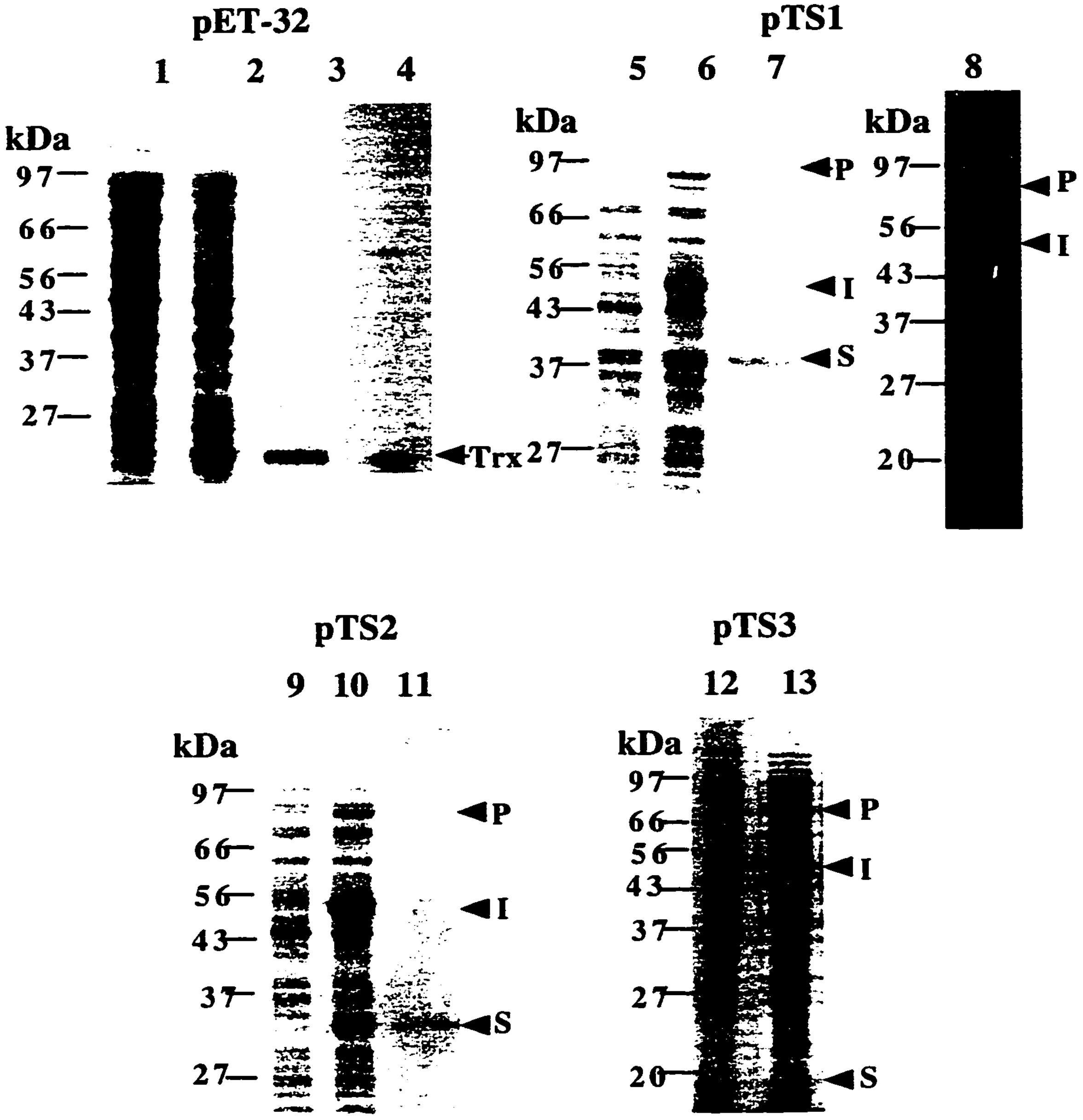

pET-32
1
2
3
4
kDa
97
66
56
43
37
27
Trx
pTS1
5
6
7
kDa
97
66
56
43
37
27
P
I
S
8
kDa
97
56
43
37
27
20
P
I
pTS2
9
10
11
kDa
97
66
56
43
37
27
P
I
S
pTS3
12
13
kDa
97
66
56
43
37
27
20
P
I
S

Figure 3-5

Figure 3-6 Overexpression of fusion proteins from plasmid pTS1 at different temperatures. All samples were electrophoresed on SDS-polyacrylamide gel and stained with Coomassie Blue. Lane 1: before induction. Lane 2: after induction at 15 °C. Lane 3: after induction at 25 °C. Lane 4: after induction at 37 °C. Letters P, I, and S mark positions of precursor protein, excised intein, and spliced protein, respectively.

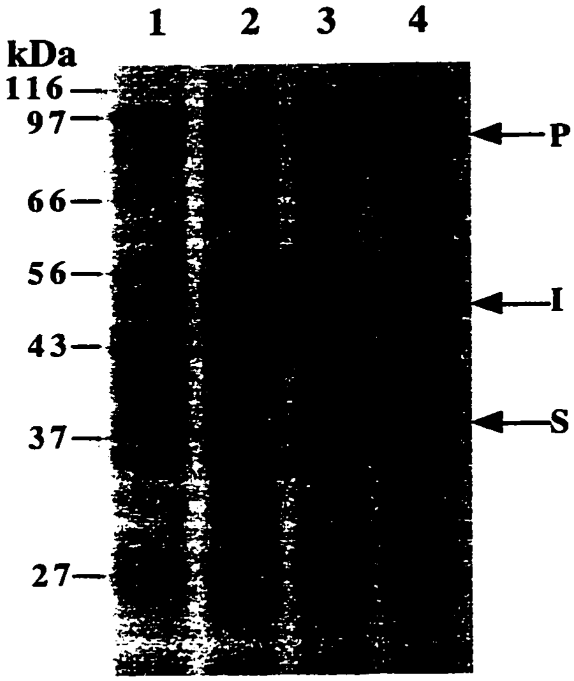


Figure 3-6

temperature is known to decrease the formation of inclusion bodies by slowing down the rate of protein synthesis. Nevertheless, the observation of an excised intein and a spliced protein in pTS1 and pTS2 indicates that the *Ssp* DnaB intein is capable of protein splicing.

3.2.3 Protein splicing of the *Ssp* DnaB intein containing deletion mutations:

A series of deletion mutations was made in the *Ssp* DnaB intein coding sequence by a nested deletion method and a PCR-based method (Fig. 3-7, see detail in sections 2.4 and 2.5.2). Deletion boundaries of five such deletion mutations are shown in Fig. 3-8. As a guide in constructing the deletion mutations, the *Ssp* DnaB intein sequence was aligned with the related but much shorter intein sequence of the *Porphyra purpurea* chloroplast DnaB protein (Fig. 3-8). Previously recognized putative intein motifs (sequence motifs A to H) were also taken into consideration. In particular, one deletion mutation (pTS1-3) was constructed to have its deleted area match closely the sequence gap between the *Ssp* DnaB intein and the *Ppu* DnaB intein.

All deletion mutations of the *Ssp* DnaB intein were placed in an expression plasmid vector in the same configuration as the control plasmid pTS1 (Fig. 3-9). For each of the fusion proteins encoded in the recombinant plasmids pTS1-1 through pTS1-5, theoretical sizes were calculated for a precursor protein, an excised intein, and a spliced protein. These recombinant plasmids were introduced into *E. coli* cells to produce corresponding fusion proteins and to observe possible protein splicing products (Fig. 3-10). The presence of protein splicing is indicated by the production of a spliced protein and an excised intein in addition to a precursor protein. In each case, the precursor protein and the spliced protein were identified by a combination of three criteria: 1) their apparent sizes matched closely the predicted sizes; 2) they were bound specifically by metal affinity resin (a property of their poly-His tag sequence); 3) they were recognized by the S protein in a Western blot analysis (a property of their S-tag sequence). It is apparent from Fig. 3-10

Figure 3-7 Illustration of strategies used to construct deletions in the *Ssp* DnaB intein. The progenitor plasmid pTS1 is a circular DNA with intein coding sequence (thick line) and vector sequence (thin line). Deletion constructs (pTS1-1 to pTS1-5) were generated from pTS1 either by a nested deletion method (top) or a PCR-based method (bottom).

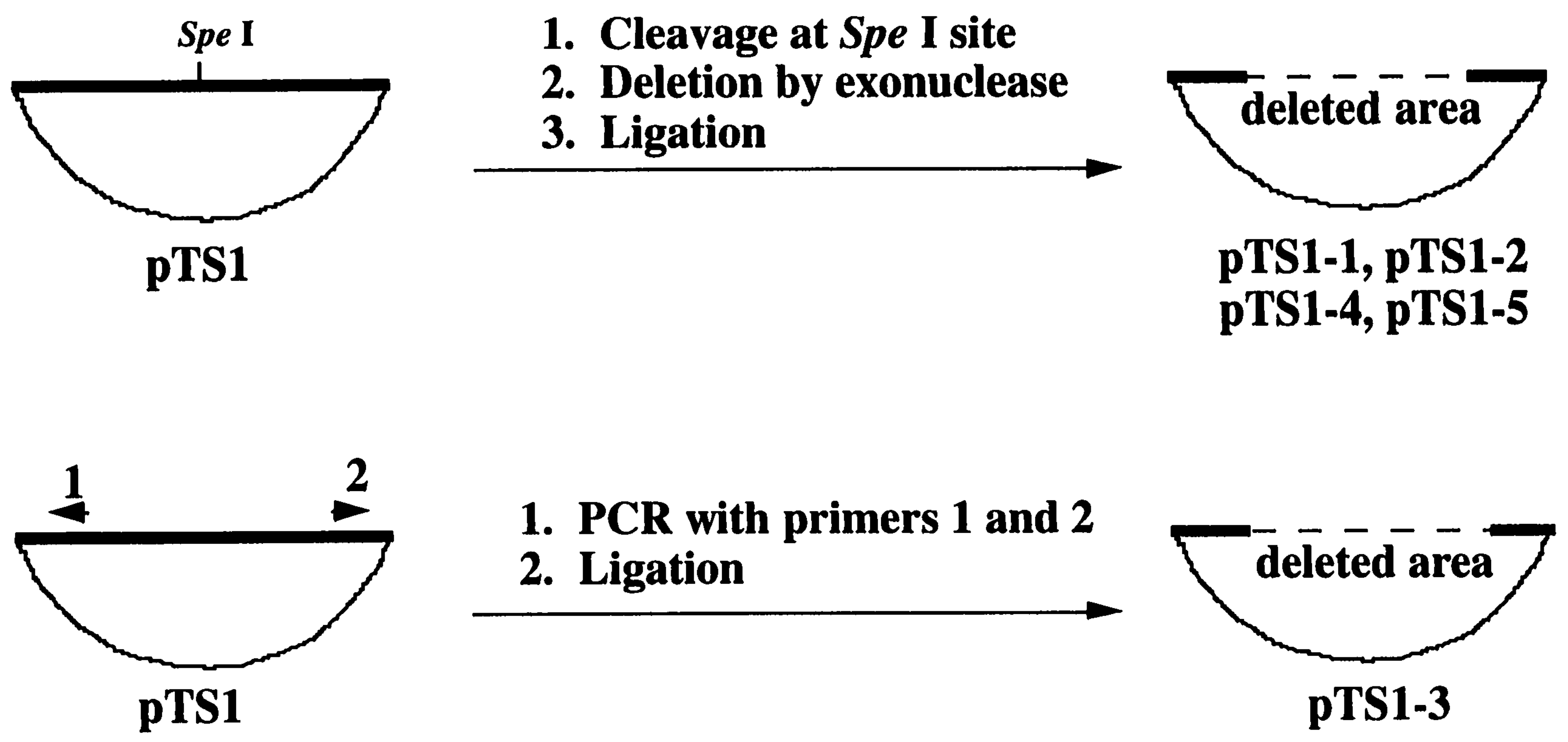
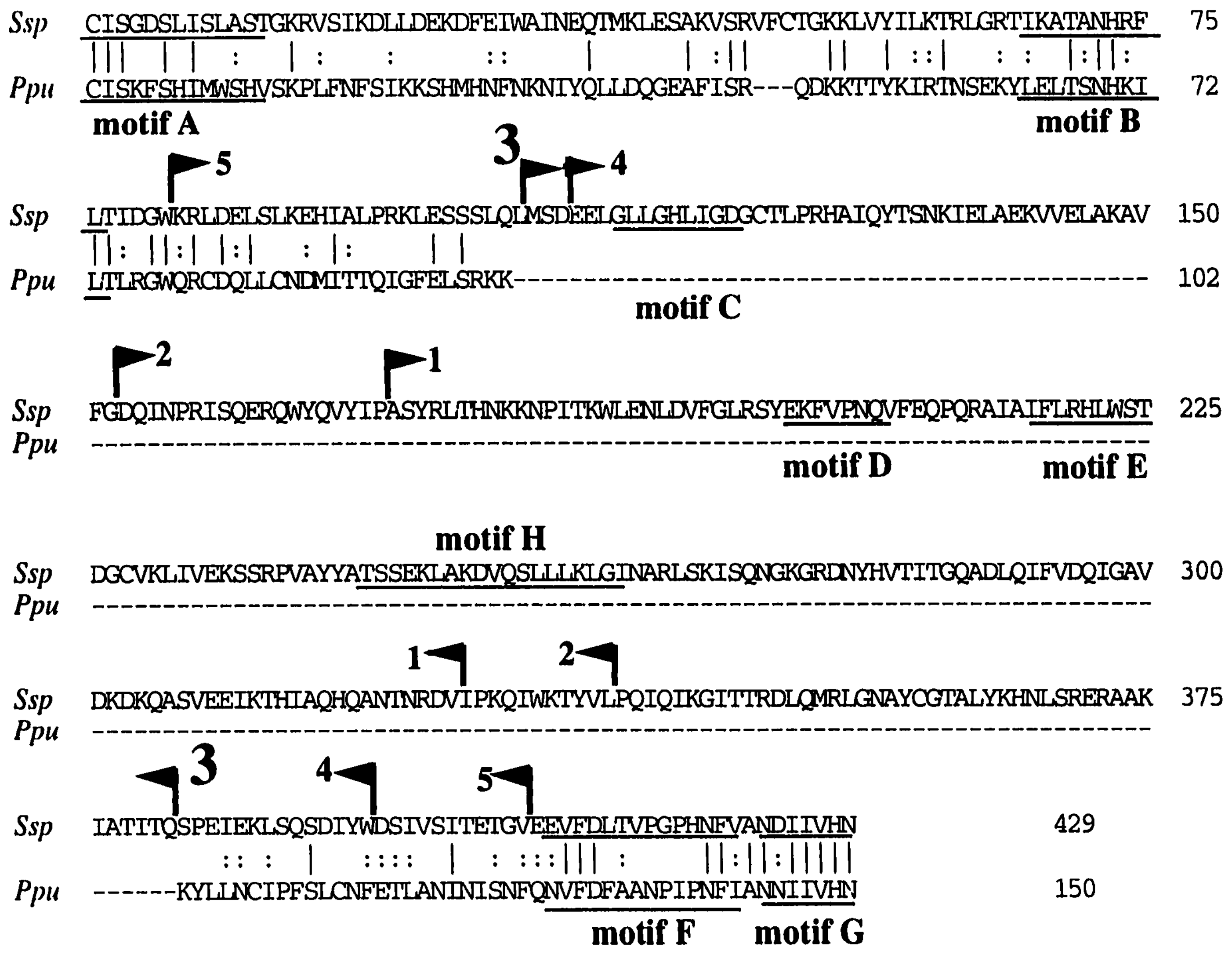

Spe I
pTS1
1. Cleavage at Spe I site
2. Deletion by exonuclease
3. Ligation
deleted area
pTS1-1, pTS1-2
pTS1-4, pTS1-5
1
2
pTS1
1. PCR with primers 1 and 2
2. Ligation
deleted area
pTS1-3

Figure 3-7

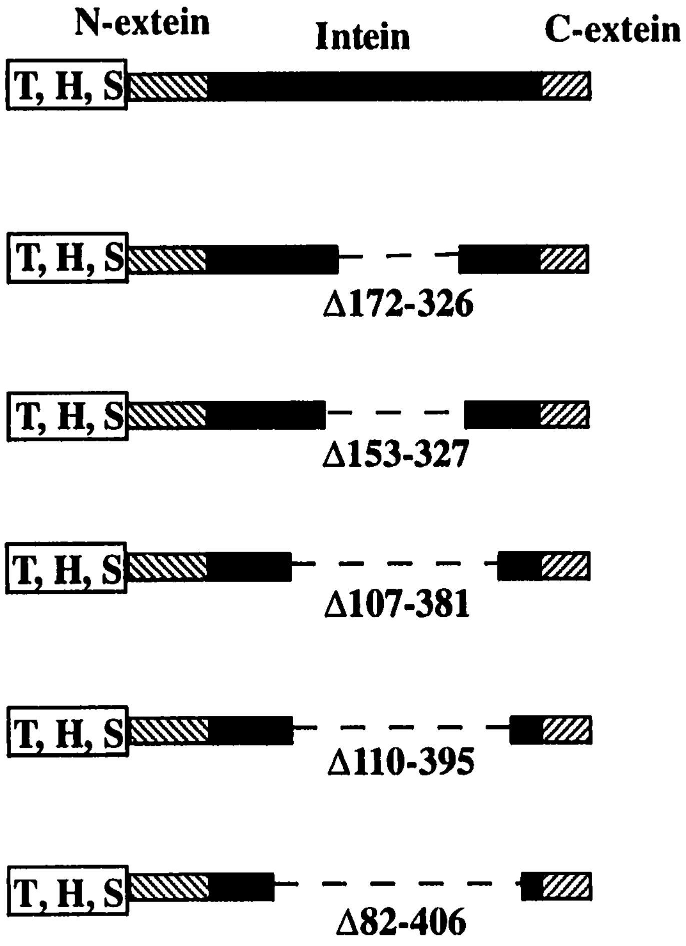
Figure 3-8 Sequences of deletion mutants of the *Ssp* DnaB intein. Protein sequence of the *Ssp* DnaB intein (*Ssp*) is shown and aligned with sequence of the DnaB intein of the *Porphyra purpurea* chloroplast (*Ppu*). Flag 1 through flag 5 mark deletion boundaries of the deletion construct pTS1-1 through pTS1-5, respectively. For example, the upstream and downstream deletion boundaries of pTS1-1 are marked by the right-pointing flag 1 and the left-pointing flag 1, respectively. Blocks A to H are conserved intein motifs (Perler *et al.*, 1997). Symbols: - represent gaps introduced to optimize the alignment; | and : mark positions of identical and similar amino acids, respectively.



```
Ssp  CISGDSLISLASTGKRVSIKDLLDEKDFEIWAINEQTMKLESAKVSRVFCTGKKLVYILKTRLGRTIKATANHRF   75
     |||  | |   :   | :   :    ::    |     | :||    ||  | ::|   : :  |:||:
Ppu  CISKFSHIMWSHVSKPLFNFSIKKSHMHNFNKNIYQLLDQGEAFISR---QDKKTTYKIRTNSEKYLELTSNHKI   72

Ssp  LTIDGWKRLDELSLKEHIALPRKLESSSLQLMSDEELGLLGHLIGDGCTLPRHAIQYTSNKIELAEKVVELAKAV  150
     ||: ||:| |:|   : |:      | |
Ppu  LTLRGWQRCDQLLCNDMITTQIGFELSRKK---------------------------------------------  102

Ssp  FGDQINPRISQERQWYQVYIPASYRLTHNKKNPITKWLENLDVFGLRSYEKFVPNQVFEQPQRAIAIFLRHLWST  225
Ppu  ---------------------------------------------------------------------------

Ssp  DGCVKLIVEKSSRPVAYYATSSEKLAKDVQSLLLKLGINARLSKISQNGKGRDNYHVTITGQADLQIFVDQIGAV  300
Ppu  ---------------------------------------------------------------------------

Ssp  DKDKQASVEEIKTHIAQHQANTNRDVIPKQIWKTYVLPQIQIKGITTRDLQMRLGNAYCGTALYKHNLSRERAAK  375
Ppu  ---------------------------------------------------------------------------

Ssp  IATITQSPEIEKLSQSDIYWDSIVSITETGVEEVFDLTVPGPHNFVANDIIVHN               429
           :: :  |   ::::  |  : :::||| :       ||:||:|||||
Ppu  ------KYLLNCIPFSLCNFETLANINISNFQNVFDFAANPIPNFIANNIIVHN               150
```

Figure 3-8

Figure 3-9 Schematic illustration of recombinant proteins encoded by the individual plasmids containing either the wild-type *Ssp* DnaB intein (pTS1) or the intein with deletions (pTS1-1 through pTS1-5). The intein (solid box), extein (hatched box), and vector-encoded (open box) sequences are indicated as in Figure 3-4. Deleted areas of the intein are marked by dashed lines, and their boundaries are specified by the numbers. For each construct, calculated molecular weights are listed for the predicted precursor protein, the excised intein and the spliced protein.



	Precursor protein	Excised intein	Spliced protein
pTS1	86 kDa	49 kDa	37 kDa
pTS1-1	69 kDa	32 kDa	37 kDa
pTS1-2	64 kDa	27 kDa	37 kDa
pTS1-3	54 kDa	17 kDa	37 kDa
pTS1-4	53 kDa	16 kDa	37 kDa
pTS1-5	49 kDa	12 kDa	37 kDa

Figure 3-9

that protein splicing occurred in cells containing recombinant plasmids pTS1-1, pTS1-2, and pTS1-3, as indicated by the production of the spliced protein in each case. In cells containing plasmid pTS1-3, an excised intein band was also identified by its size and Western blot analysis using anti-intein antibody (Fig. 3-10). This mini-intein has an expected size of 17-kDa, though it appeared at the position of about 20 kDa on the gel. However, the fact that it can be recognized by the anti-intein antiserum confirmed its identity. In cells containing plasmids pTS1-1 and pTS1-2, an excised intein was also expected because of the observation of the spliced protein. However, such a protein band was not identified due to a low level of its accumulation. In cells containing plasmids pTS1-1, pTS1-2 and pTS1-3, protein splicing did not proceed as efficiently as in cells containing the control plasmid pTS1 with the wild-type intein. Western blot intensities of the precursor and the spliced protein bands were measured to estimate the amount of spliced protein as a percentage of the total (spliced protein plus precursor protein). The ratio of spliced protein over the total was measured to be 78%, 15%, 23%, and 41% for pTS1, pTS1-1, pTS1-2, and pTS1-3, respectively (Fig. 3-10, lanes 21-24). In cells containing plasmids pTS1-4 and pTS1-5, no detectable amount of spliced protein was observed, indicating the absence of detectable protein splicing. On the more sensitive Western blot, a minor band was observed just beneath the precursor protein band for each of the deletion constructs (Fig. 3-10, lanes 22-26). The size of this minor band seems to suggest a cleavage or break down product. In conclusion, protein splicing did occur in the fusion proteins containing deletion mutants of *Ssp* DnaB intein. This observation suggests that the centrally located sequences of an intein are not required for its protein splicing activity.

3.2.4 Protein splicing of the *Ssp* DnaB intein with non-native exteins:

To study the requirement of extein sequence for protein splicing, a new recombinant plasmid, pMST, was constructed. This plasmid was derived from a

Figure 3-10 Protein splicing of the *Ssp* DnaB intein containing deletion mutations. *E. coli* cells containing specified plasmids were induced at 25 °C, and the induced proteins were analyzed by SDS-polyacrylamide gel electrophoresis followed by Coomassie Blue staining or Western blot analysis. Proteins in lanes 1-3, 4-6, 7-9, 10-11, 12-13, 14-16, and 17-19 are from plasmids pTS1, pTS1-1, pTS1-2, pTS1, pTS1-3, pTS1-4, and pTS1-5, respectively. Lanes 1, 4, 7, 10, 12, 14, and 17: before induction by IPTG. Lanes 2, 5, 8, 11, 13, 15, and 18: after induction by IPTG. Lanes 3, 6, 9, 16, and 19: proteins purified by using metal affinity resin that recognizes the poly-histidine tag. Lanes 20-26: Western blot, using the S protein that recognizes the S tag, on total proteins from cells transformed with pET32 (lane 20), pTS1 (lane 21), pTS1-1 (lane 22), pTS1-2 (lane 23), pTS1-3 (lane 24), pTS1-4 (lane 25), and pTS1-5 (lane 26). Lanes 27 and 28: Western blot using an intein-specific antiserum, on total proteins of cells containing pTS1 (lane 27) or pTS1-3 (lane 28). Letters P, I, S, and Trx mark positions of precursor protein, excised intein, spliced protein, and thioredoxin protein, respectively.

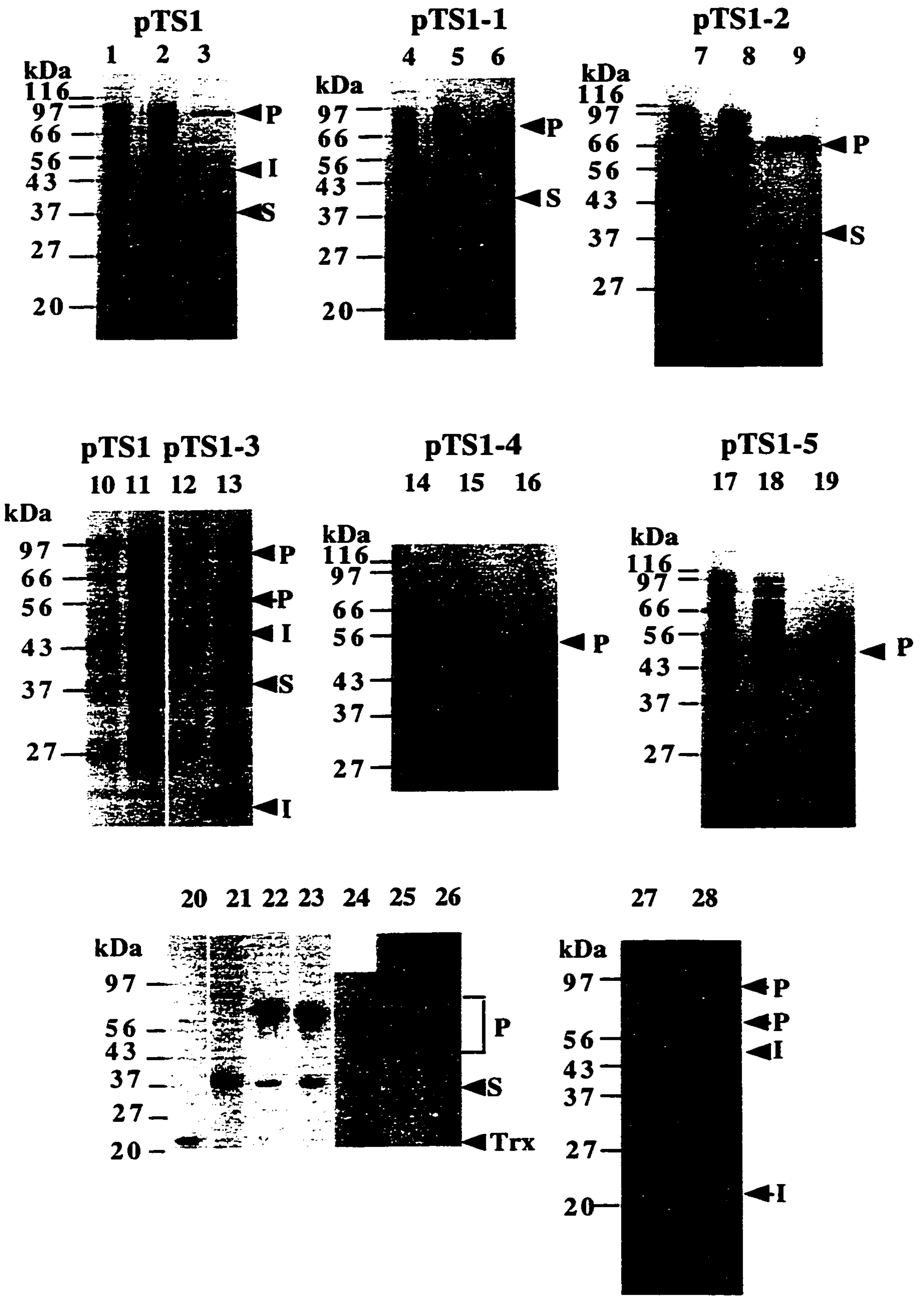
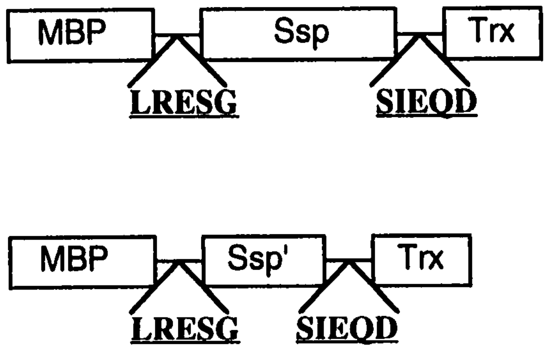

pTS1
1 2 3
kDa
116
97
66
56
43
37
27
20
P
I
S
pTS1-1
4 5 6
kDa
97
66
56
43
37
27
20
P
S
pTS1-2
7 8 9
kDa
116
97
66
56
43
37
27
P
S
pTS1 pTS1-3
10 11 12 13
kDa
97
66
56
43
37
27
P
P
I
S
I
pTS1-4
14 15 16
kDa
116
97
66
56
43
37
27
P
pTS1-5
17 18 19
kDa
116
97
66
56
43
37
27
P
20 21 22 23 24 25 26
kDa
97
56
43
37
27
20
P
S
Trx
27 28
kDa
97
56
43
37
27
20
P
P
I
I

Figure 3-10

previously constructed plasmid pMYT1 which encodes a tripartite fusion protein consisting of the *E. coli* Maltose-binding protein, Yeast *Sce* VMA intein, and *E. coli* Thioredoxin (Chong *et al.*, 1996). The *Ssp* DnaB intein coding sequence was amplified by PCR using a pair of oligonucleotide primers: 5'-TCCATGAATTCCGTCTTGTTCGATACTGTT-ATGGA and 5'-ACAAGCCTCGAGTTAAGAGAGAGTGGCTGCAT. Since pMYT1 contains restriction sites flanking the yeast intein-coding sequence, it is easy to insert this piece of sequence was inserted into pMYT1 to replace the yeast intein-coding sequence (Y) to form the new construct pMST. In this new construct, the maltose-binding protein and the thioredoxin replace the majority of the native *Ssp* DnaB exteins, leaving only five amino acid residues of native extein sequences at the two splicing junctions (Fig. 3-11). Retaining the 5 proximal native extein residues on both sides of the intein is designed to avoid potential disturbance of the intein splicing activity by proximal foreign extein residues. Similarly, plasmid pMS'T1 was constructed by replacing the yeast intein coding sequence (Y) with the 154-aa mini-intein from pTS1-3. Each of these plasmids was introduced into *E. coli* cells for expression of the fusion proteins and observation of possible protein splicing products. After three hours of induction by adding IPTG, total cellular proteins were resolved on SDS-polyacrylamide gels and stained with Coomassie Blue. In cells containing plasmid pMST, three protein products were observed (Fig. 3-12). Their sizes corresponded to the predicted size of a precursor protein (MST, 106 kDa), a spliced protein (MT, 57 kDa), and an excised intein (S, 49 kDa). Similarly, three protein products were also observed in cells containing the plasmid pMS'T1, with sizes matching closely the predicted size of a precursor protein (MS'T, 74 kDa), a spliced protein (MT, 57 kDa), and an excised mini intein (S', 17 kDa). The identities of these protein products were confirmed by Western blot analysis. The precursor proteins and the spliced proteins were specifically recognized by anti-MBP (N-extein) antibody and anti-thioredoxin (C-extein) antibody. In pMS'T1, the excised intein was specifically recognized by the anti-intein antiserum raised against the *Ssp* DnaB intein. There was so little accumulation of the

Figure 3-11 Schematic illustration of recombinant proteins encoded by plasmids containing either the wild-type *Ssp* DnaB intein (pMST) or the *Ssp* DnaB mini-intein (pMST1). In each case, the *Ssp* DnaB exteins were replaced by maltose-binding protein (MBP) and thioredoxin protein (Trx). Only 5 amino acid residues of native extein were left at each of the two splice junctions. The native extein sequences are underlined. For each construct, calculated molecular weights are listed for the predicted precursor protein, the excised intein and the spliced protein.



	Precursor (MS/S'T)	Intein (S/S')	Spliced protein (MT)
pMST	106 kDa	49 kDa	57 kDa
pMS'T1	74 kDa	17 kDa	57 kDa

Figure 3-11

Figure 3-12 Protein splicing of the *Ssp* DnaB intein and mini-intein flanked by non-native exteins. *E. coli* cells containing individual plasmids were induced at 37 °C, and the induced proteins were analyzed by SDS-polyacrylamide gel electrophoresis followed by Coomassie Blue staining or Western blot. Lanes 1 and 2: proteins from pMS'T1, before and after induction, respectively. Lanes 3 and 4: proteins from pMST, before and after induction, respectively. Lanes 5 and 6: Western blot using anti-MBP antiserum on proteins of lanes 2 and 4, respectively. Lanes 7 and 8: Western blot using anti-thioredoxin antiserum on proteins of lanes 2 and 4, respectively. Lane 9: Western blot using anti-*Ssp* DnaB intein antiserum on proteins of lane 2. MST (MS'T), MT, and S (S') mark positions of precursor protein, spliced protein, and excised intein, respectively.

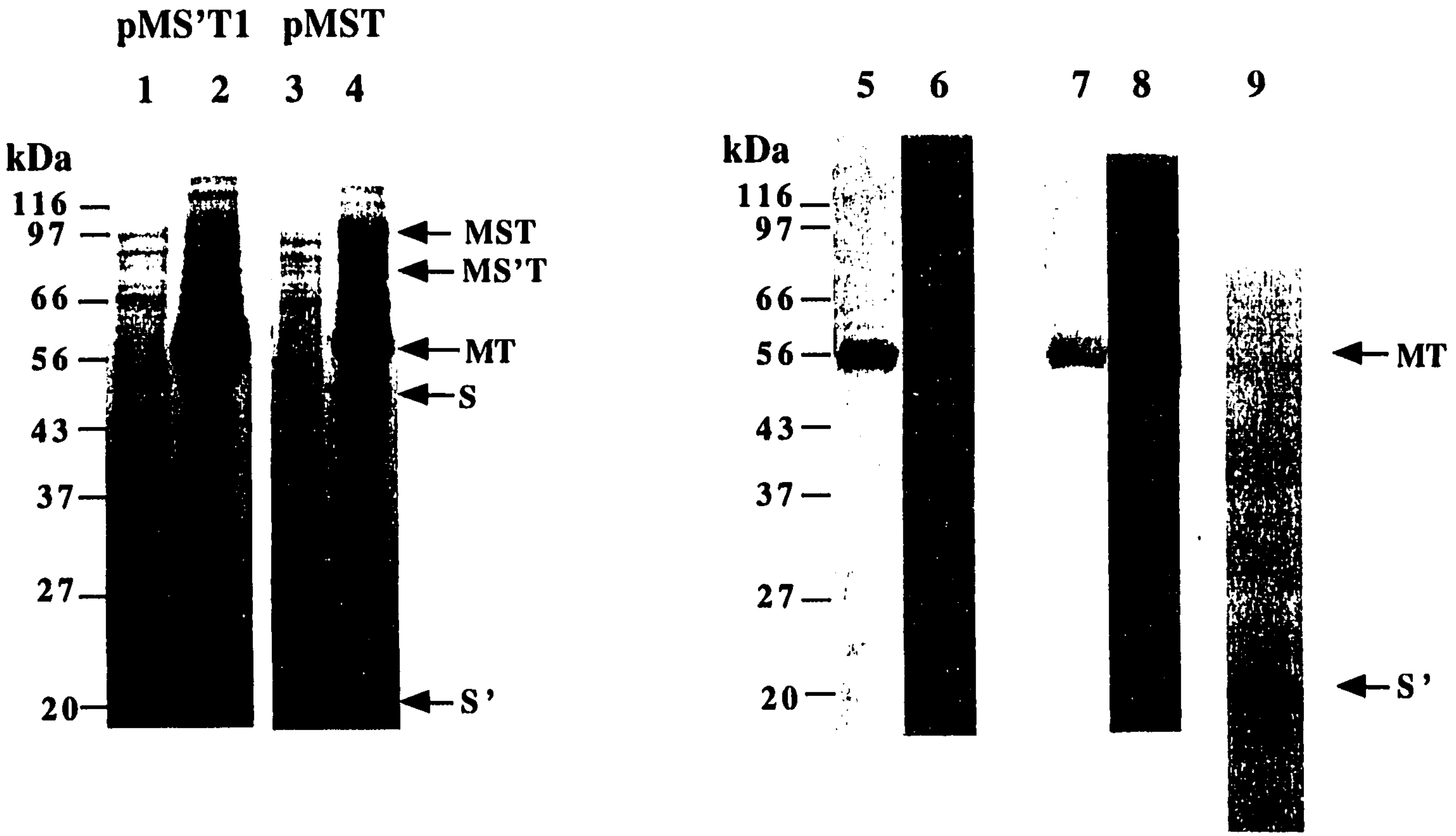

pMS'T1
pMST
1
2
3
4
kDa
116
97
66
56
43
37
27
20
MST
MS'T
MT
S
S'
5
6
7
8
9
kDa
116
97
66
56
43
37
27
20
MT
S'

Figure 3-12

precursor protein that the precursor proteins were not detected in the Western blot analysis, when the proteins were diluted 100 fold. This observation also indicates that protein splicing is more efficient for the pMS'T construct than it is for the pTS1-3 construct. The two constructs have identical intein sequences but different extein sequences. Also, pMS'T showed efficient protein splicing at 15 °C, 25 °C and 37 °C, while pTS1-3 showed little protein splicing at 37 °C (data not shown). These results indicate that the intein itself contains sufficient information for efficient protein splicing to occur.

To investigate whether the 5-aa native extein sequences at the splicing junctions are required for the intein's splicing activity, the 5 native residues at the N-terminal splice junction were replaced with sequence from the maltose-binding protein (Fig. 3-13, pMS'T6). Four of the 5 native extein residues at the C-terminal splice junction were also replaced by thioredoxin sequence, leaving the first Ser of C-extein to fulfill the requirement of nucleophilic residue at this position (Fig. 3-13, pMS'T8). The recombinant plasmids were transformed into *E. coli* cells. The fusion proteins were induced by adding IPTG. After 3 hours of induction, total cellular proteins were resolved on SDS-polyacrylamide gels and stained with Coomassie Blue (Fig. 3-14). In cells containing plasmid pMS'T6, only a protein product corresponding to the precursor (MS'T, 73 kDa) was observed, indicating the lack of splicing. In cells containing plasmid pMS'T8, three protein products were observed. Their sizes correspond to the precursor protein (73 kDa), the spliced protein (56 kDa) and the excised mini intein (17 kDa), indicating efficient protein splicing by this construct. Therefore, the substitution of the 5-aa native N-extein sequence with MBP sequence blocked the protein splicing, while the substitution of the 4-aa native C-extein sequence with thioredoxin sequence did not affect the protein splicing.

3.2.5 Protein splicing of a split *Ssp* DnaB mini-intein:

In testing for protein *trans*-splicing, plasmid pMS'T-split was constructed from

Figure 3-13 Schematic illustration of recombinant proteins encoded by plasmids containing the *Ssp* DnaB mini-intein and different native exteins. In pMS'T6, the 5-aa native N-extein sequence is replaced by maltose-binding protein sequence. In pMS'T8, 4 of the 5 amino acids of the native C-extein sequence are replaced by thioredoxin sequence. The native extein sequences are underlined. MBP, Ssp', and Trx represent maltose-binding protein, *Ssp* DnaB mini-intein, and thioredoxin protein, respectively.

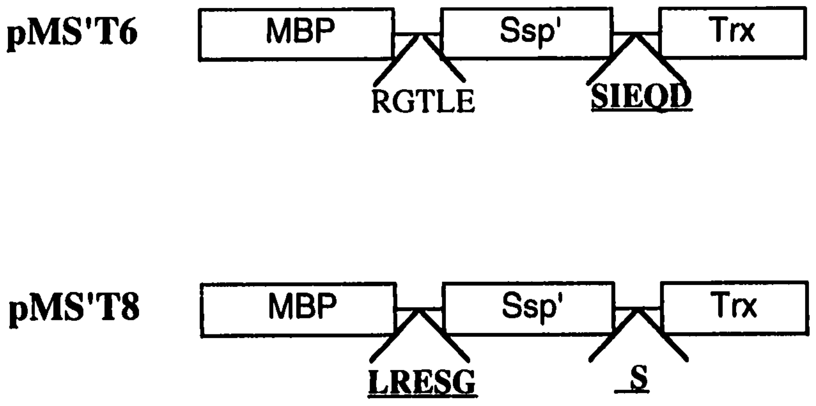


Figure 3-13

Figure 3-14 Protein splicing of the *Ssp* DnaB mini-intein in MS'T fusion proteins without the 5 aa of native extein residues . *E. coli* cells containing individual plasmids (pMS'T6 and pMS'T8) were induced at 37 °C, and the induced proteins were analyzed by SDS-polyacrylamide gel electrophoresis followed by Coomassie Blue staining. Lanes 1 and 3, before induction. Lanes 2 and 4, after induction. MS'T, MT, and S' mark positions of precursor protein, spliced protein, and excised mini-intein, respectively. The discrepancy on the migration of the proteins in lanes 1 and 3 are possibly due to the fact that the proteins were produced in different inductions and run on different gels.

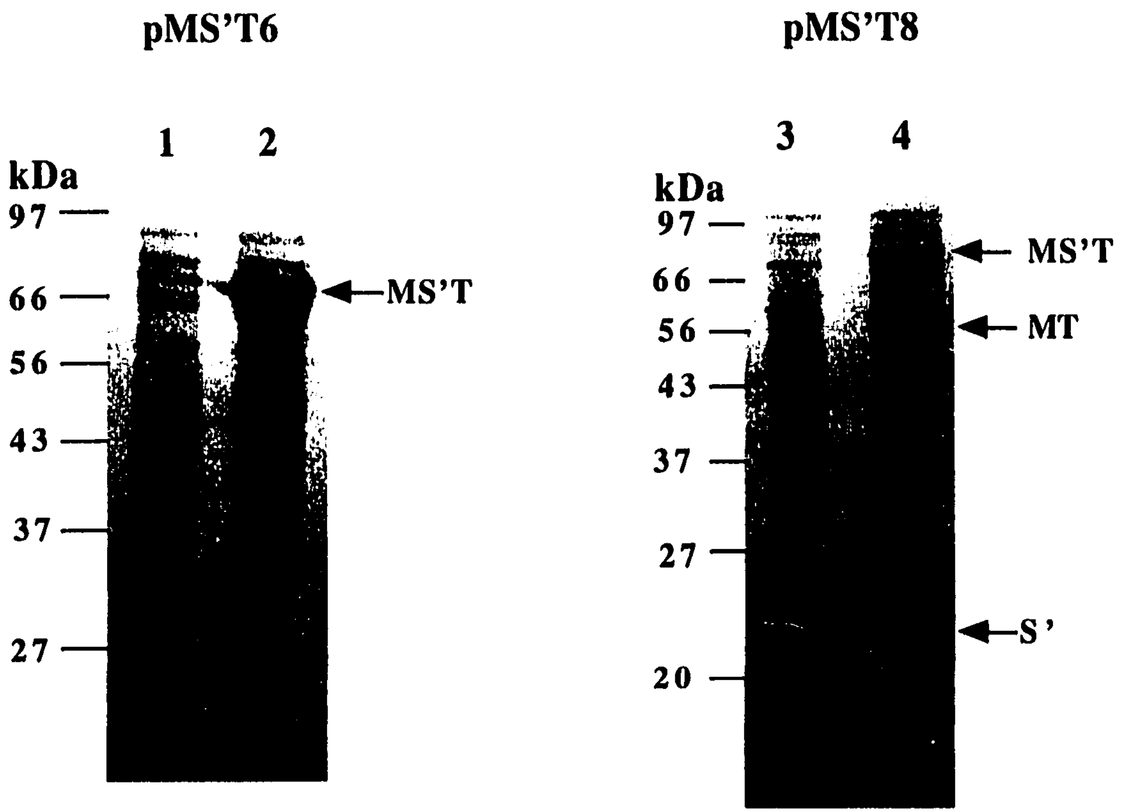

pMS'T6
1
2
kDa
97
66
56
43
37
27
MS'T
pMS'T8
3
4
kDa
97
66
56
43
37
27
20
MS'T
MT
S'

Figure 3-14

pMS'T1 by splitting the functional mini-intein in pMS'T1 into two parts (Fig. 3-15). This was achieved by inserting in the intein coding sequence a cassette consisting of [translation termination codon] - [Shine-Dalgarno sequence] - [translation initiation codon]. A PCR-mediated method was used to construct this plasmid. First, a linear DNA fragment was amplified from the circular pMS'T1 DNA in a polymerase chain reaction, using the Advantage cDNA polymerase mix and a pair of oligonucleotide primers: 5'-GGAGGTTTAAAATATGTCACCAGAAATAGAAAAGTTGTC, and 5'-CCTCATTATAATTGTAAAGAGGAGCTTTCT. The amplified linear DNA molecule was then circularized to form pMS'T-split. The resulting pMS'T-split is essentially a 2-gene operon, with the first gene (gene I) encoding the N-extein sequence plus the N-terminal sequence of the intein, and the second gene (gene II) encoding the C-terminal sequence of the intein plus the C-extein sequence. A control plasmid pMS'T-n was also constructed by inserting only a translation termination codon in the intein coding sequence, without introducing the Shine-Dalgarno sequence and the translation initiation codon (Fig. 3-15).

In *E. coli* cells containing the pMS'T-split plasmid, production of a spliced protein was observed (Fig. 3-16). This spliced protein is, by design, identical to the spliced protein produced from pMS'T1 through protein *cis*-splicing. In addition to having the expected size, the spliced protein from pMS'T-split was recognized by antiserum against maltose-binding protein and by antiserum against thioredoxin, but not by antiserum against the intein, all as expected. In addition to the spliced protein, the protein product of the first gene (gene I) of the two-gene operon also accumulated. This protein was first identified by its size, which is the same as the protein product of pMS'T-n. Also as expected, this protein was recognized by antiserum against maltose-binding protein and by antiserum against the intein, but not by antiserum against thioredoxin. A protein product of the second gene (gene II) of the two-gene operon was not observed. The accumulation of gene

Figure 3-15 Schematic illustration of fusion proteins encoded by the corresponding plasmids containing the split mini-intein and the mini-intein. In pMS'T-split, the DNA sequence of a small insertion is shown, with the termination codon TAA and the initiation codon ATG enclosed in boxes and the Shine-Dalgarno sequence underlined. In each case, the open boxes represent vector-encoded sequences. MBP and T represent maltose-binding protein and thioredoxin protein, respectively. Calculated molecular weights of the predicted protein products are listed.

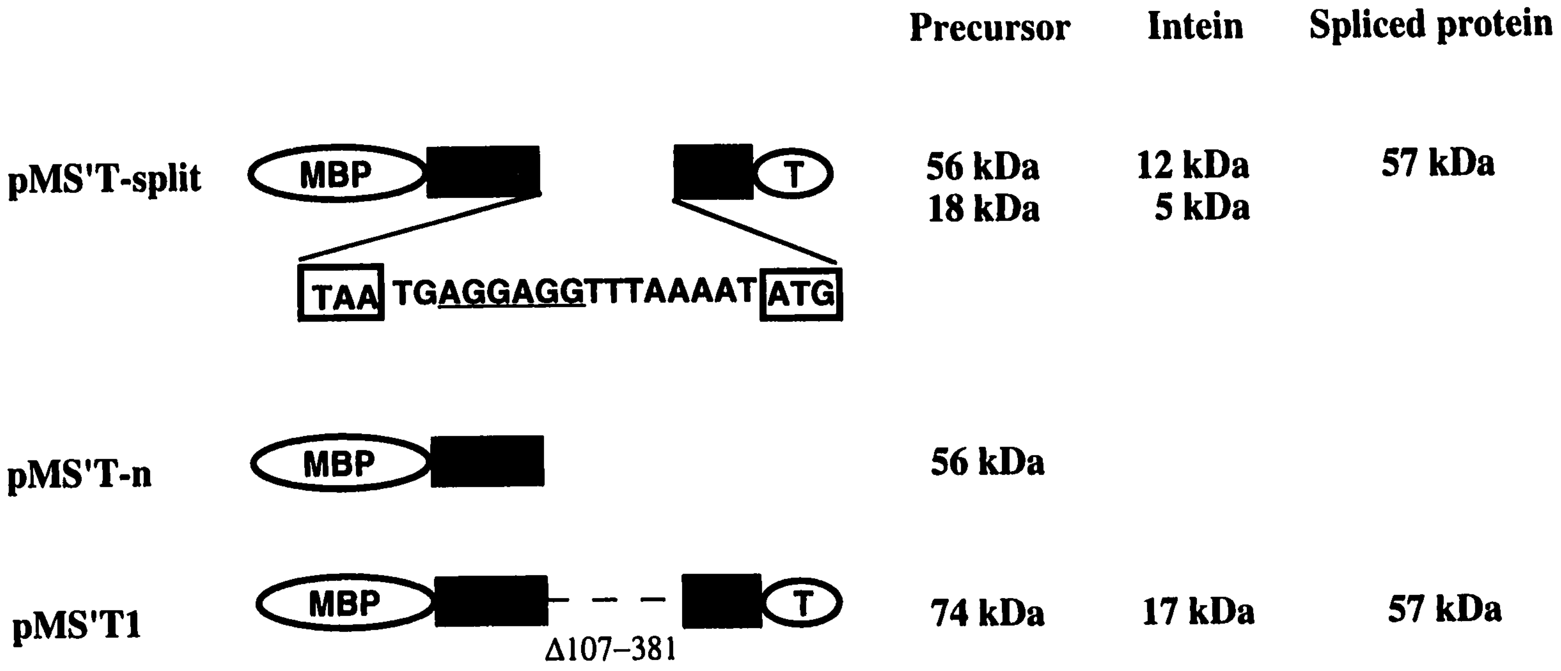

Precursor
Intein
Spliced protein
pMS'T-split
MBP
T
TAA TGAGGAGGTTTAAAAT ATG
56 kDa
18 kDa
12 kDa
5 kDa
57 kDa
pMS'T-n
MBP
56 kDa
pMS'T1
MBP
T
Δ107–381
74 kDa
17 kDa
57 kDa

Figure 3-15

Figure 3-16 Protein splicing by the split *Ssp* DnaB mini-intein. *E. coli* cells containing individual plasmids were induced at 37 °C, and the induced proteins were analyzed by SDS-polyacrylamide gel electrophoresis followed by Coomassie Blue staining or Western blotting. Lane 1: proteins of control cells (before induction, lane 1). Lanes 2-4: proteins of cells containing pMS'T1 (lane 2), pMS'T-split (lane 3), pMS'T-n (lane 4) (after induction). Lanes 5, 6, and 7: Western blot using anti-MBP antiserum on proteins of lanes 2, 3, and 4, respectively. Lanes 8, 9, and 10: Western blot using anti-thioredoxin antiserum on proteins of lanes 2, 3, and 4, respectively. Lanes 11 and 12: Western blot using anti-intein antiserum on proteins of lanes 3 and 2, respectively. Letters P, S, and I mark positions of precursor protein (MS'T), spliced protein (MT), and excised mini-intein (S'), respectively. Letter N marks protein product of pMS'T-n.

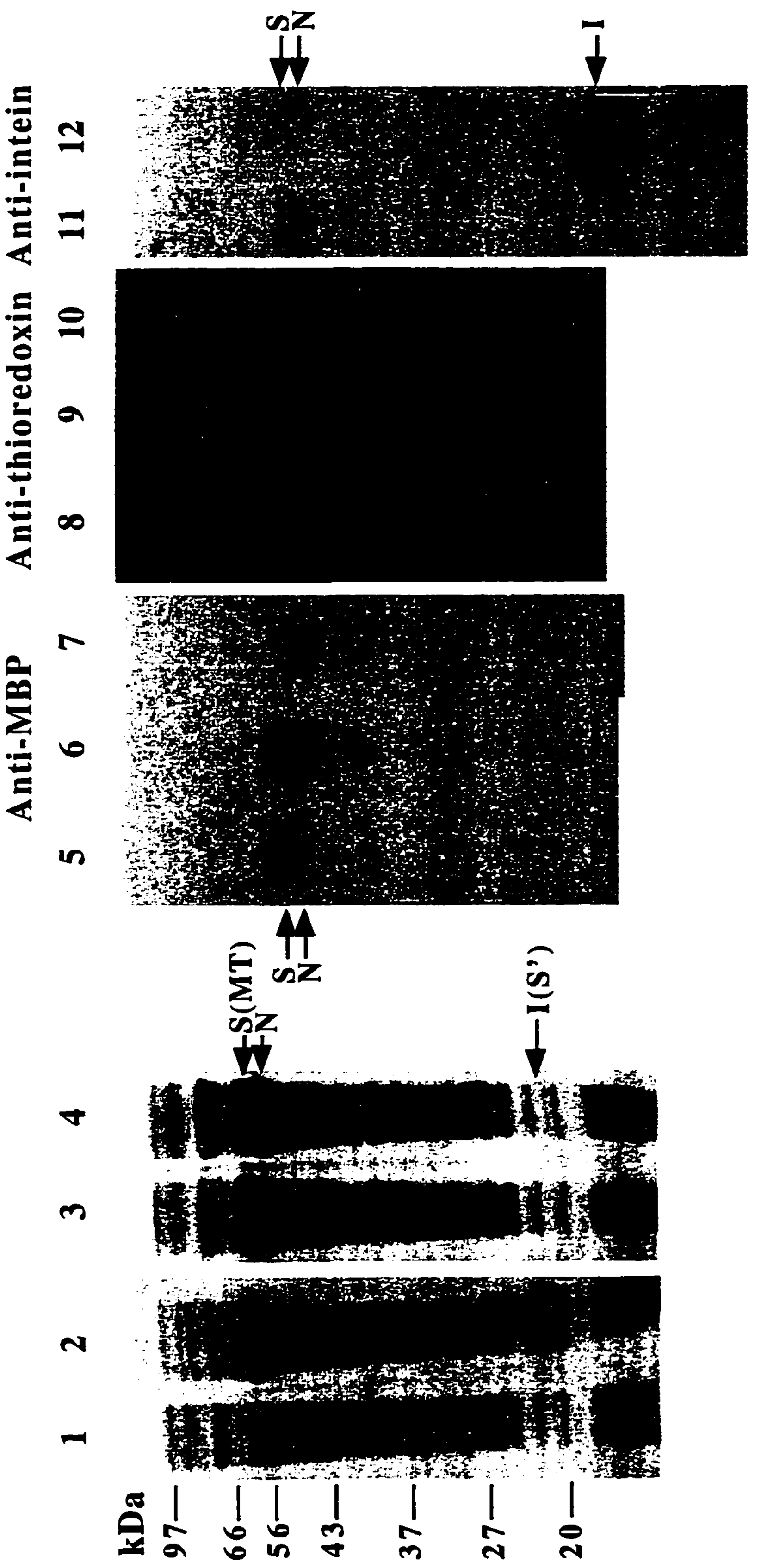

kDa
97
66
56
43
37
27
20
1
2
3
4
S(MT)
N
I(S')
Anti-MBP
5
6
7
S
N
Anti-thioredoxin
8
9
10
Anti-intein
11
12
S
N
I

Figure 3-16

I protein but not gene II protein indicates that gene I protein was produced in molar excess (relative to gene II protein), probably due to a less than 100% translational coupling between gene I and gene II. Two excised intein fragments, predicted to be 12 kDa and 5 kDa, respectively, were not observed, most likely due to their small sizes, weak recognition by the anti-intein antiserum that was raised against a continuous full-length *Ssp* DnaB intein, and/or rapid degradation in the *E. coli* cells. However, the observation of a spliced protein (MT) in cells containing pMS'T-split indicate the protein *trans*-splicing occurred.

3.2.6 Mutagenesis study of the *Ssp* DnaB mini-intein:

Previous studies on the mechanism of protein splicing revealed a requirement for the conserved residues at the splice junctions (Davis *et al.*, 1992; Hodges *et al.*, 1992; Cooper *et al.*, 1993; Hirata *et al.*, 1992; Xu and Perler, 1996). It is still not clear whether these residues are the only critical residues required for splicing. In searching for other amino acid residues required for protein splicing, mutations were introduced into the *Ssp* DnaB mini-intein. The recombinant plasmids used in this study were constructed from pMS'T1 by cassette replacement and PCR-mediated random mutagenesis method. Each of these plasmids was introduced into *E. coli* cells for protein expression. In pMS'T3, the first nucleophilic residue (Cys) of the *Ssp* DnaB intein was changed to Ser. In *E. coli* cells containing this construct, two proteins were produced: the precursor protein (MS'T, 74 kDa) and the C-terminal cleavage product (MS', 61 kDa) (Fig. 3-17). Another C-terminal cleavage product, T, was not detected, possibly due to its small size (13 kDa). Since no spliced protein was detected, the Cys1Ser mutation in MS'T blocked the protein splicing reaction.

To investigate the effect of the internal residues on intein's splicing activity, mutations were introduced randomly throughout the entire *Ssp* DnaB mini-intein (S') in the

Figure 3-17 Protein splicing of MS'T with the Cys1Ser mutation. *E. coli* cells containing the plasmid were induced at 37 °C, and the induced proteins were analyzed by SDS-polyacrylamide gel electrophoresis, followed by Coomassie Blue staining or Western blot analysis. Lanes 1 and 2: proteins from cells containing control plasmid pMS'T1, before and after induction, respectively. Lanes 3 and 4: proteins from cells containing pMS'T3 (Cys1Ser), before and after induction, respectively. Lanes 5 and 6: Western blot using anti-MBP antiserum on proteins of lanes 2 and 4. Lanes 7 and 8: Western blot using anti-thioredoxin antiserum on proteins of lanes 2 and 4. Letters MS'T, MT, MS' and S' represent precursor, spliced protein, C-terminal cleavage product, and excised mini intein, respectively.

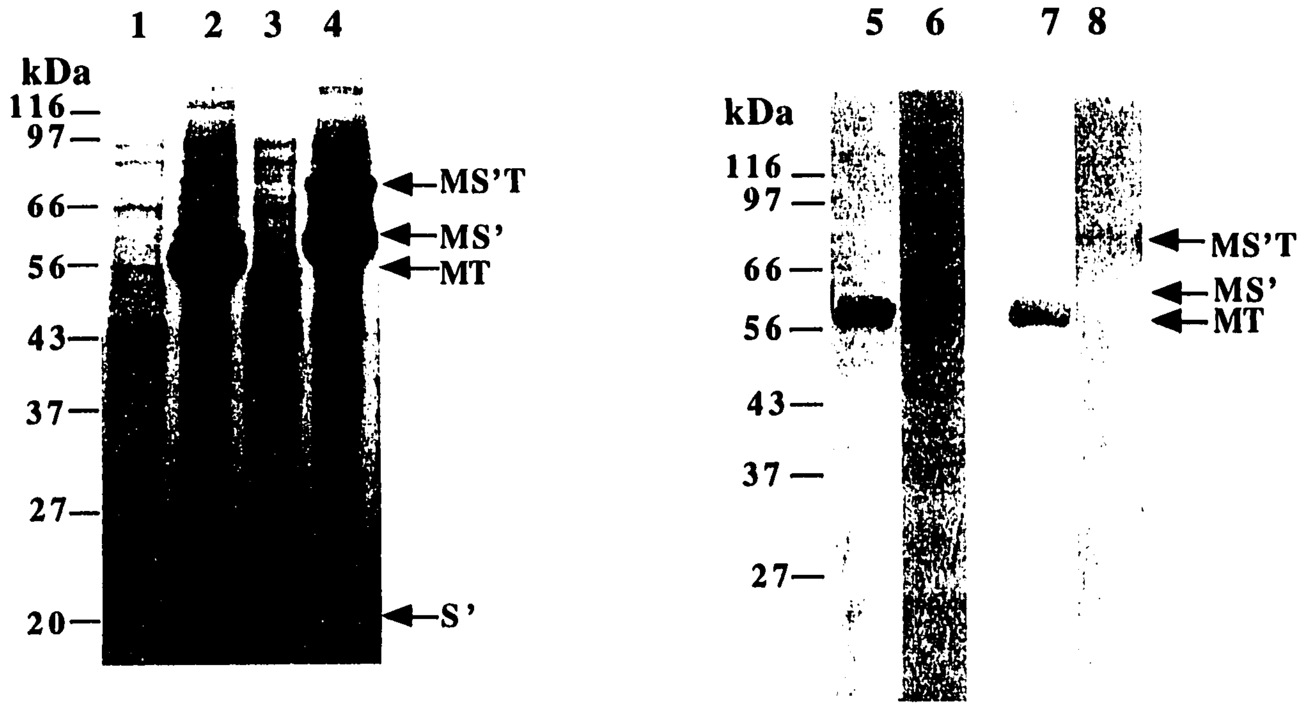

1
2
3
4
kDa
116—
97—
66—
56—
43—
37—
27—
20—
MS'T
MS'
MT
S'
5
6
7
8
kDa
116
97
66
56
43
37
27
MS'T
MS'
MT

Figure 3-17

MST construct by error-prone PCR. The mutagenized 0.5-kbp *Xho*I-*Eco*RI fragment covers the entire *Ssp* DnaB mini-intein region. After 3-h induction at 37 °C, total proteins of *E. coli* cells were examined by SDS-PAGE followed by Coomassie Blue staining. The splicing activity of mutants was estimated by comparing the density of the 74-kDa precursor band (MST) with the density of the 57-kDa band of spliced protein (MT). Eight mutants (out of 100 putative mutants) accumulated the precursor and did not show detectable splicing products on the Coomassie Blue-stained gel. Plasmid DNAs of these defective mutants were sequenced in the mutagenized region. The results of DNA sequencing are shown in Table 3-1. The mutations were distributed throughout the entire mutagenized region. The number of amino acid substitutions in each mutant varied from 1 to 4. Three of the eight mutants contain single amino acid substitutions. One of the three mutations, I150V, is close to the C-terminal splice junction, falling into the conserved sequence motif G (see Fig. 3-8). The other two substitutions, G80D and I126T, did not fall into any of the sequence motifs but were close to sequence motifs B and F, respectively (see Fig. 3-8). These results revealed that some internal residues of the intein are also required for its protein splicing activity.

3.2.7 Endonuclease activity assay of the *Ssp* DnaB intein:

To study the endonuclease activity of *Ssp* DnaB intein *in vitro*, this intein was purified by using the IMPACT™ protein purification system. The *Ssp* DnaB intein coding sequence was inserted into expression vector TYB163 at *Nde*I and *Eco*RI sites. In the resulting plasmid, pTSYB, the *Ssp* DnaB intein coding sequence was placed in front of the *Sce* VMA intein coding sequence and the chitin binding domain coding sequence (Fig. 3-18). This construct produced a tripartite fusion protein of 104 kDa containing the *Ssp* DnaB intein - Yeast intein - chitin binding domain. To block possible *in vivo* cleavage of the *Ssp* DnaB intein at the C terminus of the intein, the last Asn of the *Ssp* DnaB intein was changed to Ala by using an oligo primer containing such a mutation in the polymerase chain

Table 3-1 Mutations that abolish protein splicing of MS'T1 in *E. coli*

Mutant No.	No. of Amino Acid Changes	Mutations
1	2	N72K, L99P
3	1	G80D
30	2	K15N, L40P
31	1	I126T
42	2	E86D, E132K
66	1	I150V
87	4	G141R, P142T, N148H, I150T
100	4	K68N, V134G, N144T, I151F

* The amino acids are numbered to their position in the *Ssp* DnaB mini-intein. Residues 1-106 in the mini-intein correspond to resides 1-106 in the wild-type intein. Residues 107-154 in the mini-intein correspond to residues 381-429 in the wild-type intein (see Fig. 3-8).

reaction. After induction, the crude extract was loaded onto a chitin column. The fusion protein was bound onto the chitin column while other proteins were washed away. The *Ssp* intein was then eluted from the column after adding DTT to induce cleavage at the *Ssp* DnaB intein - yeast intein junction (Fig. 3-18). Plasmid pLSH2 which contains the homing site for the *Ssp* DnaB intein was first linearized by restriction enzyme *Sca*I. It was then incubated with the purified *Ssp* DnaB intein in buffers containing different concentrations of NaCl and Mg^{2+}, so that the digestion with *Ssp* DnaB intein should yield two fragments of 0.76 and 2.0 kbp. However, after 2 hours of digestion at 37 °C in different buffers, only the original linear plasmid was detected (Fig. 3-19).

A more sensitive method was then used to investigate the endonuclease activity of the *Ssp* DnaB intein *in vivo*. The *Ssp* DnaB intein homing site was cloned into pTS1 at an *Eco*RI site. In this construct (pHC-194), the homing site was placed downstream of the *Ssp dnaB* gene (Fig. 3-20). The expression of the *dnaB* gene would lead to production of the *Ssp* DnaB intein and linearization of pHC-194 at the homing site, if the intein is an active endonuclease and cuts at this homing site. After induction, the total DNA was extracted and digested with *Pst*I, followed by Southern blot analysis. The digestion by *Ssp* DnaB intein should yield fragments of 1.3 and 6.4 kbp. However, Southern blot analysis using the linear pHC-194 as probe on the total DNAs extracted at different time points detected only one band corresponding to the *Pst*I- linearized pHC-194 (Fig. 3-20). The failure in detection of any digested products may suggest that there is no detectable endonuclease activity for this intein under such experimental conditions.

Figure 3-18 Purification of the *Ssp* DnaB intein using chitin affinity chromatography. **A**. Schematic illustration of purification of *Ssp* DnaB intein by chitin affinity chromatography. Bold letters S, Y, and CBD represent the *Ssp* DnaB intein, the *Sce* VMA intein, and chitin-binding domain, respectively. Letters C, H, and A represent the Cys1, His428 and Ala429 residues in the *Ssp* DnaB intein, respectively. **B**. Overexpression and purification of *Ssp* DnaB intein. *E. coli* cells containing pTSYB were induced at 15 °C. Proteins were resolved by SDS-polyacrylamide gel electrophoresis, followed by Coomassie Blue staining. Lane 1, before induction. Lane 2, after induction. Lane 3, soluble proteins. Lane 4, flow-through from chitin column. Lane 5, purified *Ssp* DnaB intein. SYB and S represent precursor protein and the *Ssp* DnaB intein, respectively.

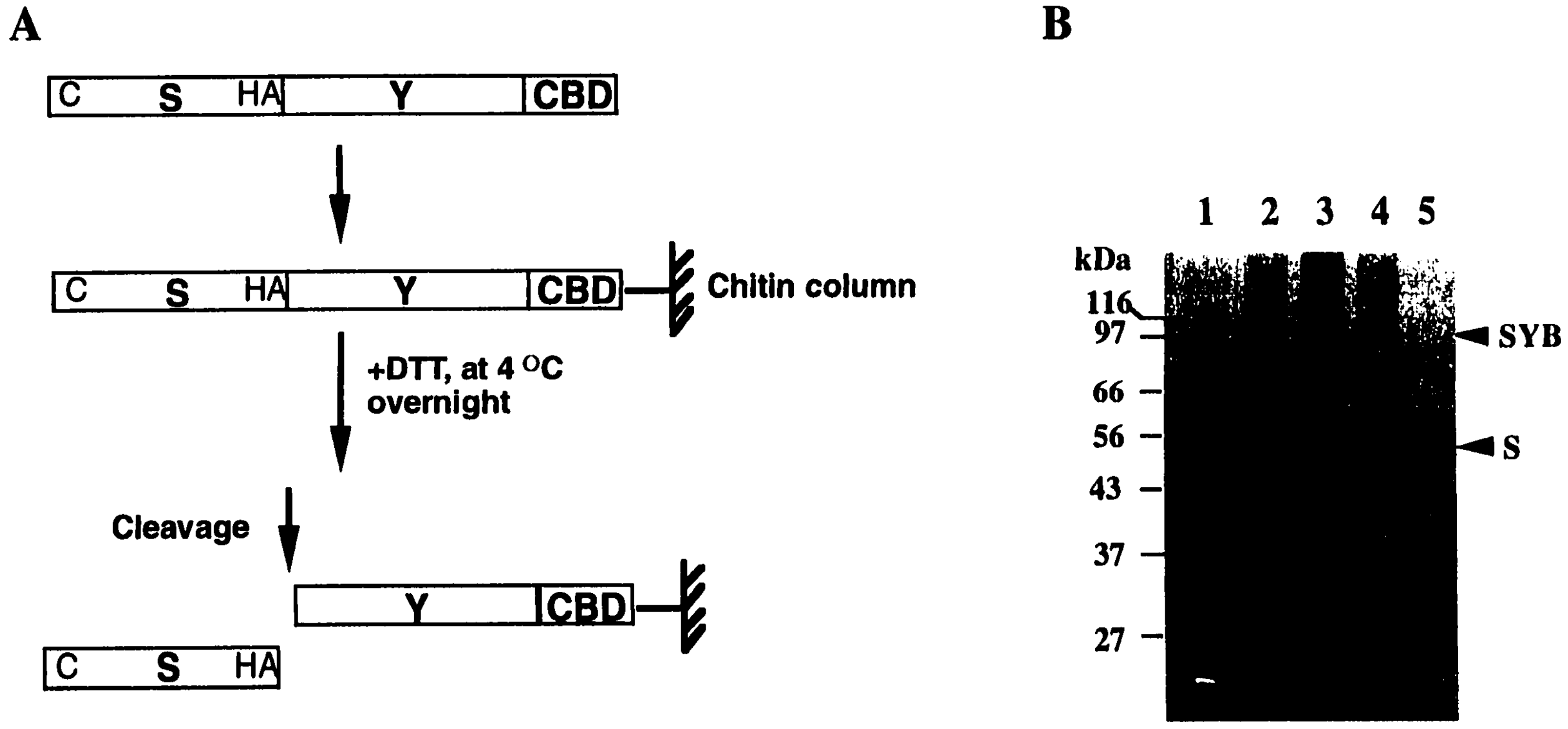

A
C
S
HA
Y
CBD
Chitin column
+DTT, at 4 °C
overnight
Cleavage
B
1 2 3 4 5
kDa
116
97
66
56
43
37
27
SYB
S

Figure 3-18

Figure 3-19 *In vitro* endonuclease assay of the *Ssp* DnaB intein. **A**. Schematic illustration of construction of the substrate for the *Ssp* DnaB intein. **B**. Digestion of pLSH2 by *Ssp* DnaB intein in different buffers. Digestion was carried out at 37 °C for 2 hours. Proteinase K was then added to the digests and incubated at 37 °C for 30 min before the samples were loaded on gel. Lane 1: control (no enzyme). Lanes 2 to 6: digestions in buffers I to V (see section 2.12.2).

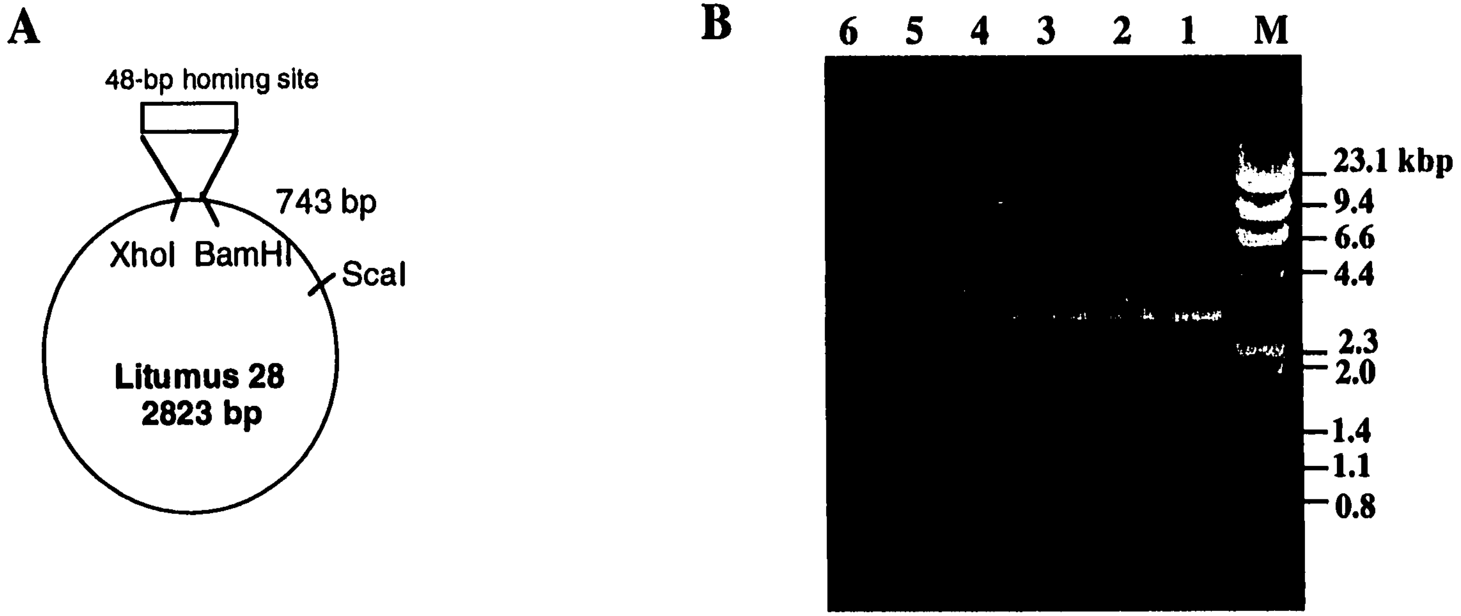

A
48-bp homing site
743 bp
XhoI
BamHI
ScaI
Litumus 28
2823 bp
B
6
5
4
3
2
1
M
23.1 kbp
9.4
6.6
4.4
2.3
2.0
1.4
1.1
0.8

Figure 3-19

Figure 3-20 *In vivo* endonuclease assay of the *Ssp* DnaB intein. **A**. Schematic illustration of the recombinant plasmid containing both the *Ssp dnaB* gene and the homing site for the *Ssp* DnaB intein. **B**. *In vivo* endonuclease assay. *E. coli* cells containing the recombinant plasmid pHC-194 were induced at 25 °C. 1-ml samples were taken after 10 min, 30 min, 1 hour, 2 hours, and 3 hours of induction. Total DNAs were extracted by the gentle lysis method and digested by *Pst*I before loading on the 1% agarose gel. After Southern transfer, the membrane was hybridized with the labeled linear plasmid pHC-194. As a control, *E. coli* cells containing the same plasmid were also incubated under the same conditions but without induction by IPTG. One ml sample was taken at the same time points as above, and the total DNAs were extracted and subjected to Southern blot.

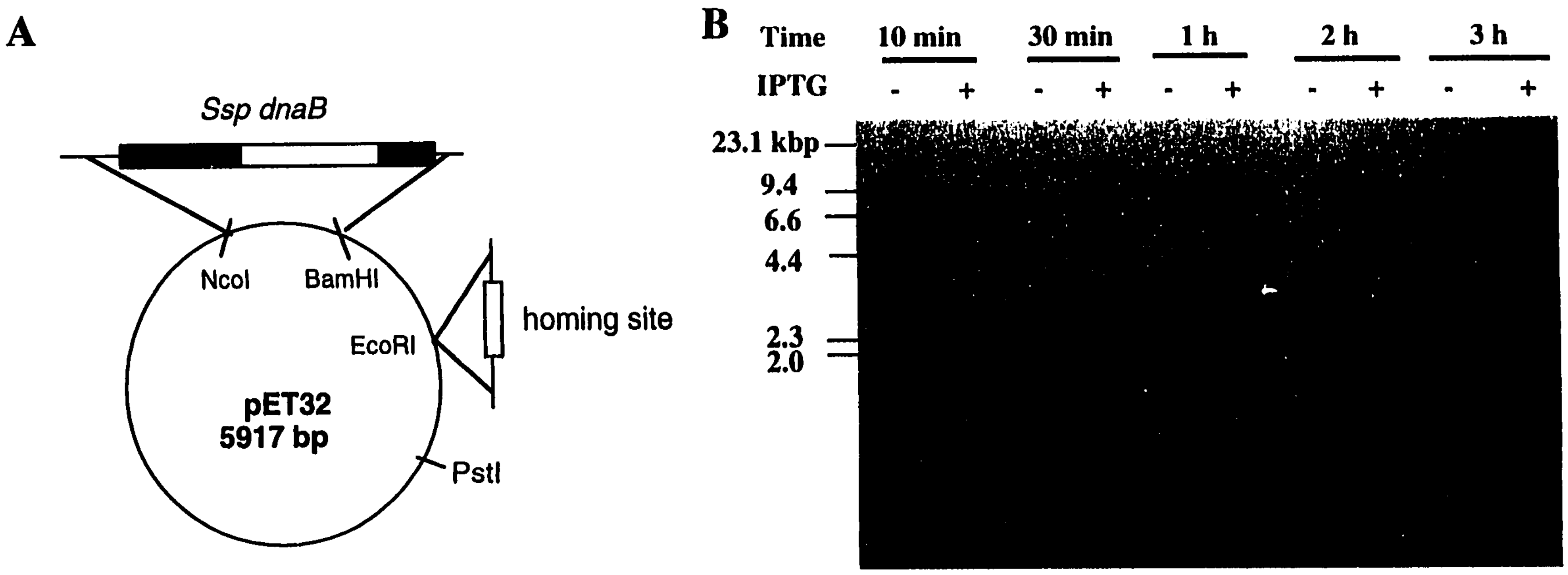

A
Ssp dnaB
NcoI
BamHI
EcoRI
homing site
pET32
5917 bp
PstI
B
Time
10 min
30 min
1 h
2 h
3 h
IPTG
- + - + - + - + - +
23.1 kbp
9.4
6.6
4.4
2.3
2.0

Figure 3-20

3.3 Discussion:

3.3.1 The DnaB protein in *Synechocystis* sp. PCC6803 contains a functional intein:

Sequence analysis of the DnaB protein of *Synechocystis* sp. PCC6803 clearly revealed the presence of a 429-aa intein. The large intervening sequence in the *Ssp dnaB* gene is an in-frame insertion in the *Ssp dnaB* gene, which is in agreement with the criteria for intein identification (Perler *et al.*, 1997). The *Ssp* DnaB intein interrupts a highly conserved region in the DnaB protein. Like the other known inteins, this intein contains the two dodecapeptide motifs. It also contains the other intein-defining features, including the splice junction residues that are critical for the splicing reaction and the other intein sequence motifs. More importantly, this intein exhibited protein splicing activity when it was produced in *E. coli* cells. Interestingly, in the *E. coli* expression system the protein splicing efficiency of this intein changes as the induction temperature changes. The splicing was more efficient at lower temperatures (15-25 °C), but less efficient at higher temperature (37 °C). The lower temperature seems to allow productive folding of the precursor protein for splicing. The formation of inclusion bodies at higher induction temperature could cause misfolding and precipitation of the precursor protein. It is not clear whether the *Ssp* DnaB protein undergoes protein splicing in the cyanobacterial cells themselves. Western blot analysis using antibody against *E. coli* DnaB protein did not detect a DnaB protein in the cyanobacterial total protein (data not shown). Although the *Ssp* DnaB protein and the *E. coli* DnaB protein share 36% sequence identity, the antibody raised against *E. coli* DnaB protein might not recognize the *Ssp* DnaB protein. Nevertheless, the *Ssp* DnaB intein is expected to be functional *in vivo*, because it clearly is so in *E. coli* cells.

Synechocystis sp. PCC6803 is one of the few eubacterial organisms in which inteins are found. Four inteins have been found in this organism so far, including the *Ssp*

DnaB intein, the *Ssp* DnaX intein in the τ subunit of DNA polymerase III (Liu and Hu, 1997a), the *Ssp* GyrB intein in a DNA gyrase B subunit (Dalgaard *et al.*, 1997a; Pietrokovski, 1998), and the split *Ssp* DnaE intein in the α subunit of DNA polymerase III (Wu *et al.*, 1998, also see Chapter IV of this thesis). The four *Ssp* inteins are not specifically related in sequence, which may suggest independent origins of these inteins. Interestingly, all of the four intein-containing proteins interact directly with DNA. This situation again highlights the observation that inteins reside predominantly in proteins of nucleic acid metabolism. Inteins are mobile genetic elements at the DNA level, by the virtue of endonucleases encoded by themselves. The host genes might be in regions of distinctive DNA or nucleoid structure. If so, the easy access of these region to invasive elements could possibly lead to the high frequency presence of inteins in these loci. On the other hand, the intein insertion sites appeared to be at or close to regions functionally important, i.e., they are close to residues involved in catalysis or substrate or cofactor binding (Dalgaard *et al.*, 1997a). Positioning in such a region of a protein could help to maintain the intein, because this kind of site may not be able to tolerate the loss of inteins, since their loss may interrupt the expression of the host protein, leading to lethal events to the host cell. When these mobile elements invade other sites by endonuclease-mediated mobility, they would tend to reenter sites of perfect excision. Invasion at imperfect excision sites could abolish the gene function.

3.3.2 Effect of extein sequences on the splicing activity of the *Ssp* DnaB intein:

The *Ssp* DnaB intein not only undergoes protein splicing when produced in *E. coli* cells, it also did so without its complete native extein sequences. In the pST constructs, the partial DnaB protein can support protein splicing. In the MST fusion protein, the native exteins were replaced by the unrelated maltose binding protein and thioredoxin, leaving only 5 aa of proximal native residues at both splice junctions, and the intein still can actively splice. This finding is consistent with previous observations that certain intein sequences are sufficient for protein splicing when placed within a foreign protein immediately before a nucleophilic residue (Davis *et al.*, 1992; Xu *et al.*, 1993; Cooper *et al.*, 1993). In these studies, the partial or whole native extein sequences of other inteins were replaced by unrelated protein sequences with a nucleophilic residue as the first reidue of C-exteins. In those constructs, efficient protein splicing was also observed. The fact that the *Ssp* DnaB intein was still active in the MST fusion proteins indicates that the intein contains sufficient information for catalyzing splicing reactions. However, since the catalytic residues of protein splicing are located at the intein termini, the proximal extein sequences can also affect splicing when the intein is placed in a foreign context. Although replacing of the C-terminal native extein sequence did not affect splicing of the *Ssp* DnaB mini-intein, when the 5 amino acid residues of native sequence at the N-terminal splicing junction (LRESG) were replaced by non-native sequence (RGTLE), only the accumulated precursor protein was detected, indicating that protein splicing was blocked. Although substitution of the Gly -1 of the *Sce* VMA intein with some other amino acids affects the protein splicing (Chong *et al.*, 1998), it is likely that the residue proximal to the N-terminal splice junction is not the only extein residue which affects protein splicing. The other adjacent extein residues may also be involved. Together, they may help in formation of the active core for protein splicing by interacting with other residues. The residues at the C-terminal junction (either extein or intein residues) are possible candidates, because the catalytic core would require the proximal position of the two splice junctions. It has been recently proposed that splicing of the *Sce* VMA intein involves interactions between the intein residues upstream of the C-terminal splice junction and the proximal extein residues upstream of the N-terminal splice junction (Nogami *et al.*, 1997). This proposal could also be applied to the *Ssp* DnaB intein. The functional involvement of the extein sequence in the splicing reaction suggests possible preference of this intein for its homing site. The intein-coding sequence would tend to insert into DNA regions encoding a peptide containing the information that prefers protein splicing. Without the restriction of such a peptide sequence, the homing events could lead to splicing-deficient mutants, which could be lethal to the host cells.

3.3.3 Defining the splicing domain in the *Ssp* DnaB intein:

Protein splicing activity and endonuclease activity are the two features of inteins. Most of the known inteins have sizes ranging from 335 to 548 amino acids. They have 8 conserved sequence motifs (motifs A to H), including two motifs (motif C and E) that are characteristics of homing endonucleases. Questions that have been raised include whether these two endonuclease motifs are involved in protein splicing, and whether the N- and C-terminal regions of inteins contain sufficient information for protein splicing. It has been shown that protein splicing is independent of endonuclease function, because an archaeal intein with a mutation that abolishes the endonuclease activity can still splice efficiently (Hodges *et al.*, 1992). It has also been shown that protein splicing is blocked by a small deletion in an intein of the Vent DNA polymerase of *Thermococcus litoralis* (Hodges *et al.*, 1992) and by a larger deletion in the central region of an intein of *Mycobacterium tuberculosis* RecA protein (Davis *et al.*, 1992). In the *Sce* VMA intein, a 7-aa deletion in the middle of the intein does not affect splicing, but larger in-frame deletions have been shown to block splicing (Cooper *et al.*, 1993). In this study, the *Ssp* DnaB intein has been shown to be capable of splicing without the central region. In the deletion construct pTS1-3, the centrally located 275 aa of the intein sequence is removed, creating a functional miniature intein that is only 154 aa in size. The middle two-thirds of *Ssp* DnaB intein sequence spans sequence motifs C, D, E, and H. The 154-aa functional mini-intein contains sequence motifs A, B, F, and G. Although the splicing efficiency of this miniature intein is approximately 50% lower than that of the wild-type intein, the fact that it can splice indicates that: (i) all the essential reacting groups that participate in the catalysis of the splicing reaction are located in the mini-intein; (ii) the mini-intein carries sufficient structural information for proper folding of the intein to bring the two splice junctions in

Figure 3-21 Functional mini-intein. A model of the *Ssp* DnaB intein is shown to relate its function with its structure. Structural domain [In, Ic] contains the N-terminal region (In, residues 1-107, approximate) and the C-terminal region (Ic, residues 381-429, approximate) of the intein sequence, while domain II represents the endonuclease domain in the middle part of the intein sequence. Domain [In, Ic] is a functional splicing domain that is sufficient for protein splicing, as was demonstrated by the deletion constructs pTS1-3 and pMS'T1.

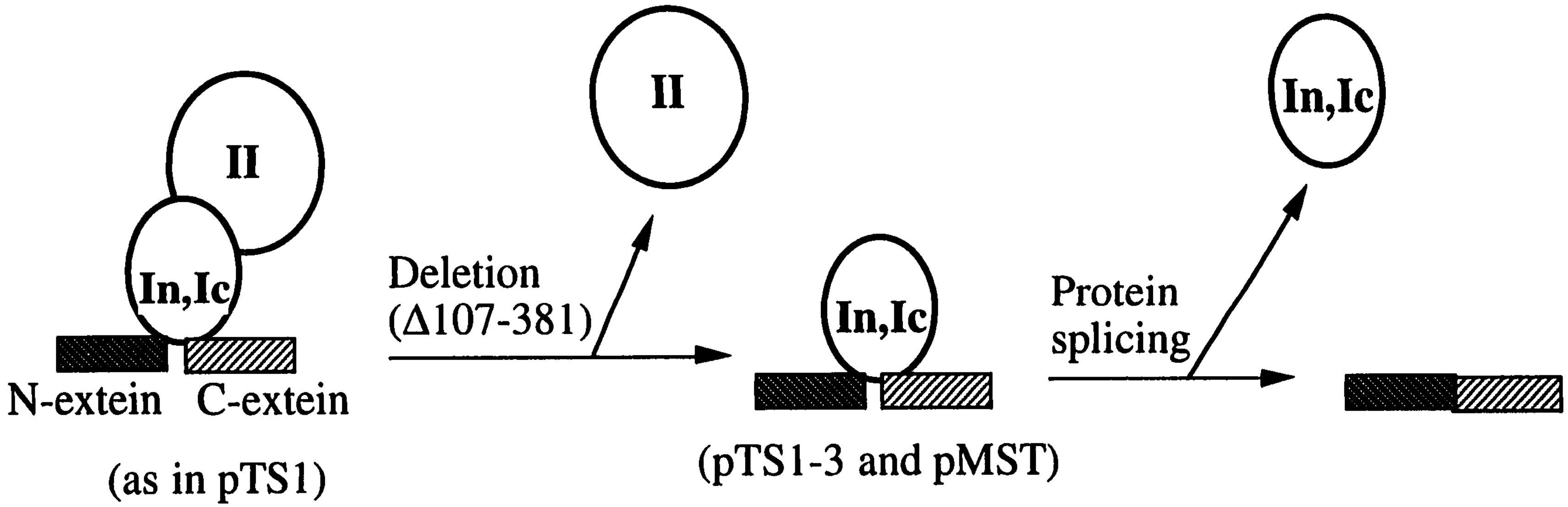

II
In,Ic
N-extein
C-extein
(as in pTS1)
Deletion
(Δ107-381)
II
In,Ic
(pTS1-3 and pMST)
Protein
splicing
In,Ic

Figure 3-21

close proximity and to precisely align all the catalytic groups.

Data presented here led to the proposal of a model describing the structure and function of the *Ssp* DnaB intein (Fig. 3-21). The model predicts two structural domains in the intein: a protein-splicing domain (domain I) that is not only sufficient for protein splicing but also structurally separable from other parts of the intein; and an endonuclease domain (domain II) that is not required for protein splicing. This model is also consistent with the recently reported crystal structure of the *Sce* VMA intein that exhibits a two-domain (domain I and II) structure (Duan *et al.*, 1997) and the crystal structure of the *Mxe* GyrA intein which lacks the endonuclease domain (Klabunde *et al.*, 1998). Other studies including mutagenesis studies (Kawasaki *et al.*, 1997) and sequence statistical modeling also support such a bipartite structure (Dalgaard *et al.*, 1997a, 1997b).

The proposed splicing domain corresponds approximately to the functional miniature intein constructed in this study (construct pTS1-3). It is formed by a coming together of the N-terminal sequence (domain In) and the C-terminal sequence (domain Ic) of the intein. These terminal sequences contain the conserved intein sequence motifs A, B, F, and G. Previous studies have shown that these terminal sequences contain several residues involved in the catalysis of protein splicing, including the Cys residue at the beginning of motif A and the Asn residue at the end of motif G. The results from this study show that these terminal sequences contain all the essential elements for protein splicing. In fact, the miniature splicing domain could be further reduced in size. A 142-aa mini-intein has been constructed by deleting 12 more resides in the central region. This mini-intein still kept its splicing activity (data not shown), indicating it still contains sufficient information for splicing. In the deletion constructs pTS1-4 and pTS1-5, further deletions into the splicing domain resulted in loss of splicing activity, suggesting a removal of essential sequences. Whether these non-functional inteins can be restored to splicing by replacing the deleted sequences with properly designed flexible linker sequences is not known. The suggested endonuclease domain (domain II) contains conserved intein sequence motifs C, D, E, and H. Motifs C and E are present in a large number of endonucleases associated with inteins and introns, and they contain acidic residues known to be important for the endonuclease activity (Gimble and Stephens, 1995). Motif H was recognized more recently as a putative intein motif (Perler *et al.*, 1997), but no clear function has been assigned to it. The results from this study show that motif H is not required for splicing, suggesting that it more likely participates in the endonuclease function.

Although a crystal structure is not available for the *Ssp* DnaB intein, structural information can be inferred from the known crystal structures of the *Sce* VMA intein and the *Mxe* GyrA intein. Statistical modeling has produced sequence alignments among the *Ssp* DnaB intein, *Sce* VMA intein, and *Mxe* GyrA intein along with many other inteins (Dalgaard *et al.*, 1997), suggesting a structural resemblance among different inteins. Based on the sequence alignments, the functional mini-intein (154 aa) derived from the *Ssp* DnaB intein corresponds to a major part (approximately 70%) of domain I (splicing domain) of the *Sce* VMA intein, while the 275-aa sequence that was deleted from the *Ssp* DnaB intein corresponds to the entire domain II (endonuclease domain) plus a part of domain I of the *Sce* VMA intein (Fig. 3-22). The crystal structure of the *Mxe* GyrA intein shows a β-core formed by the N- and C-terminal sequences of the intein, with the middle part of the intein sequence forming a disordered region and two α helices that extend from the β-core (Klabunde *et al.*, 1998). Based on the crystal structure and a sequence alignment between the *Ssp* DnaB intein and the *Mxe* GyrA intein (Dalgaard *et al.*, 1997), the functional mini-intein (154 aa) derived from the *Ssp* DnaB intein corresponds to the entire β-core of the *Mxe* GyrA intein, while lacking most of the disordered region and α helices present in the *Mxe* GyrA intein.

Figure 3-22 Comparison between the *Ssp* DnaB intein and the *Sce* VMA intein. **(A)** Alignment of the *Ssp* DnaB intein sequence with the sequence of the *Sce* VMA intein. Motifs A to H are putative intein motifs that are recognized previuosly (Perler *et al.*, 1997). The boxed area corresponds to the domain II endonuclease domain of the *Sce* VMA intein (Duan *et al.*, 1997). Two flags mark boundaries of deletion in the mutant *Ssp* DnaB intein construct pTS1-3. Symbols: -represent gaps introduced to optimize the alignment; | and : mark positions of identical and similar amino acids, respectively. **(B)** Schematic illustration of the *Ssp* DnaB intein and its deleted form (pTS1-3), compared to the *Sce* VMA intein. Boxes A to H represent the conserved intein motifs. Structural domains assigned to the *Sce* VMA intein are based on the crystal structure of this intein. In the *Ssp* DnaB intein, a putative domain II corresponds to the domain II of the *Sce* VMA intein, while domains In and Ic make up a putative domain I. The deletion mutant pTS1-3 has only domains In and Ic.

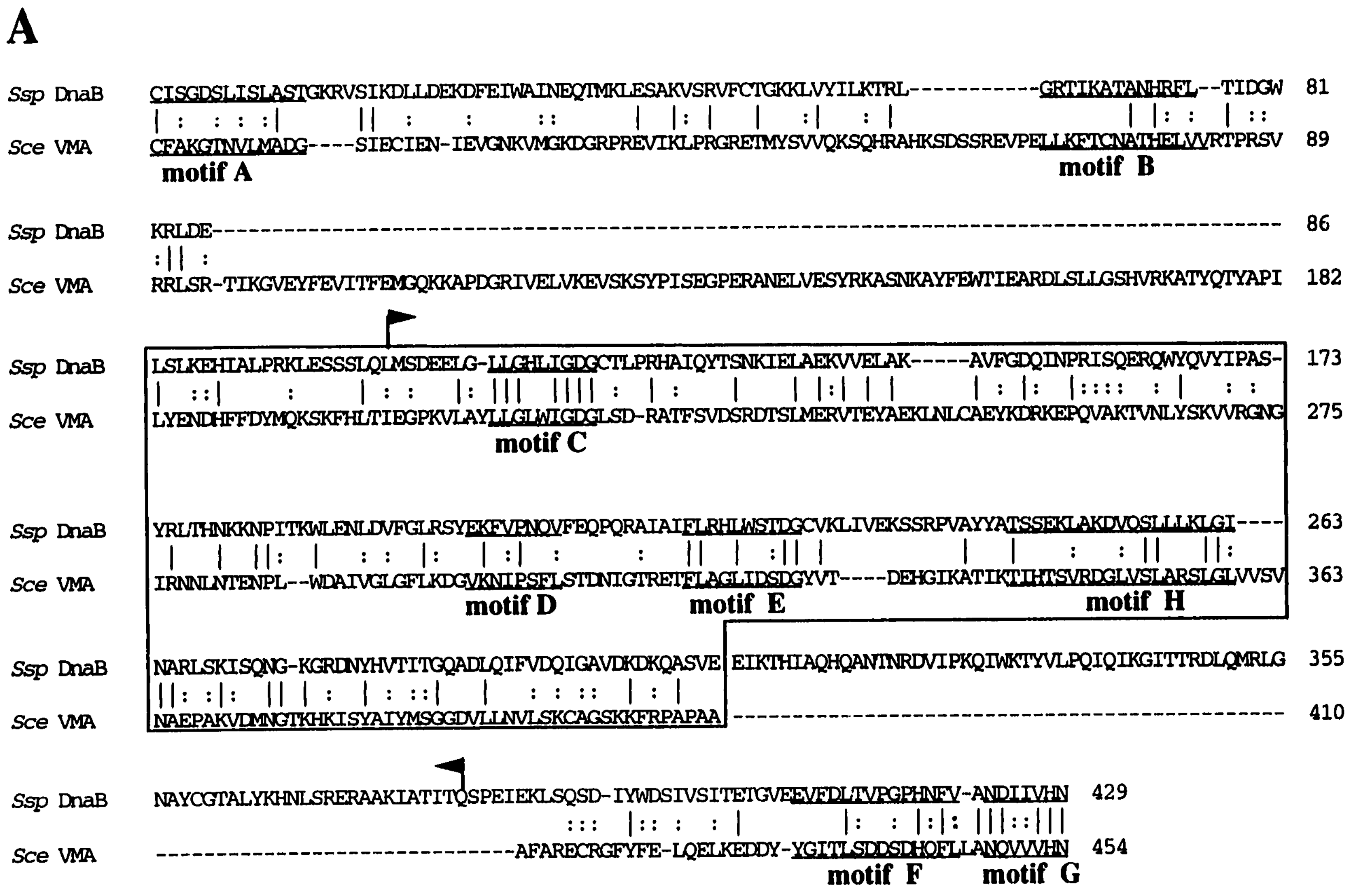

A
Ssp DnaB
CISGDSLISLASTGKRVSIKDLLDEKDFEIWAINEQTMKLESAKVSRVFCTGKKLVYILKTRL------------GRTIKATANHRFL--TIDGW
81
Sce VMA
CFAKGTNVLMADG----SIECIEN-IEVGNKVMGKDGRPREVIKLPRGRETMYSVVQKSQHRAHKSDSSREVPELLKFTCNATHELVVRTPRSV
89
motif A
motif B
Ssp DnaB
KRLDE--
86
Sce VMA
RRLSR-TIKGVEYFEVITFEMGQKKAPDGRIVELVKEVSKSYPISEGPERANELVESYRKASNKAYFEWTIEARDLSLLGSHVRKATYQTYAPI
182
Ssp DnaB
LSLKEHIALPRKLESSSLQLMSDEELG-LLGHLIGDGCTLPRHAIQYTSNKIELAEKVVELAK-----AVFGDQINPRISQERQWYQVVYIPAS-
173
Sce VMA
LYENDHFFDYMQKSKFHLTIEGPKVLAYLLGLWIGDGLSD-RATFSVDSRDTSLMERVTEYAEKLNLCAEYKDRKEPQVAKTVNLYSKVVRGNG
275
motif C
Ssp DnaB
YRLTHNKKNPITKWLENLDVFGLRSYEKFVPNQVFEQPQRAIAIFLRHLWSTDGCVKLIVEKSSRPVAYYATSSEKLAKDVQSLLLKLGI----
263
Sce VMA
IRNNLNTENPL--WDAIVGLGFLKDGVKNIPSFLSTDNIGTRETFLAGLIDSDGYVT----DEHGIKATIKTIHTSVRDGLVSLARSLGLVVSV
363
motif D
motif E
motif H
Ssp DnaB
NARLSKISQNG-KGRDNYHVTITGQADLQIFVDQIGAVDKDKQASVE EIKTHIAQHQANTNRDVIPKQIWKTYVLPQIQIKGITTRDLQMRLG
355
Sce VMA
NAEPAKVDMNGTKHKISYAIYMSGGDVLLNVLSKCAGSKKFRPAPAA --
410
Ssp DnaB
NAYCGTALYKHNLSRERAAKIATITQSPEIEKLSQSD-IYWDSIVSITETGVEEVFDLTVPGPHNFV-ANDIIVHN
429
Sce VMA
-----------------------------AFARECRGFYFE-LQELKEDDY-YGITLSDDSDHQFLLANQVVVHN
454
motif F
motif G

Figure 3-22

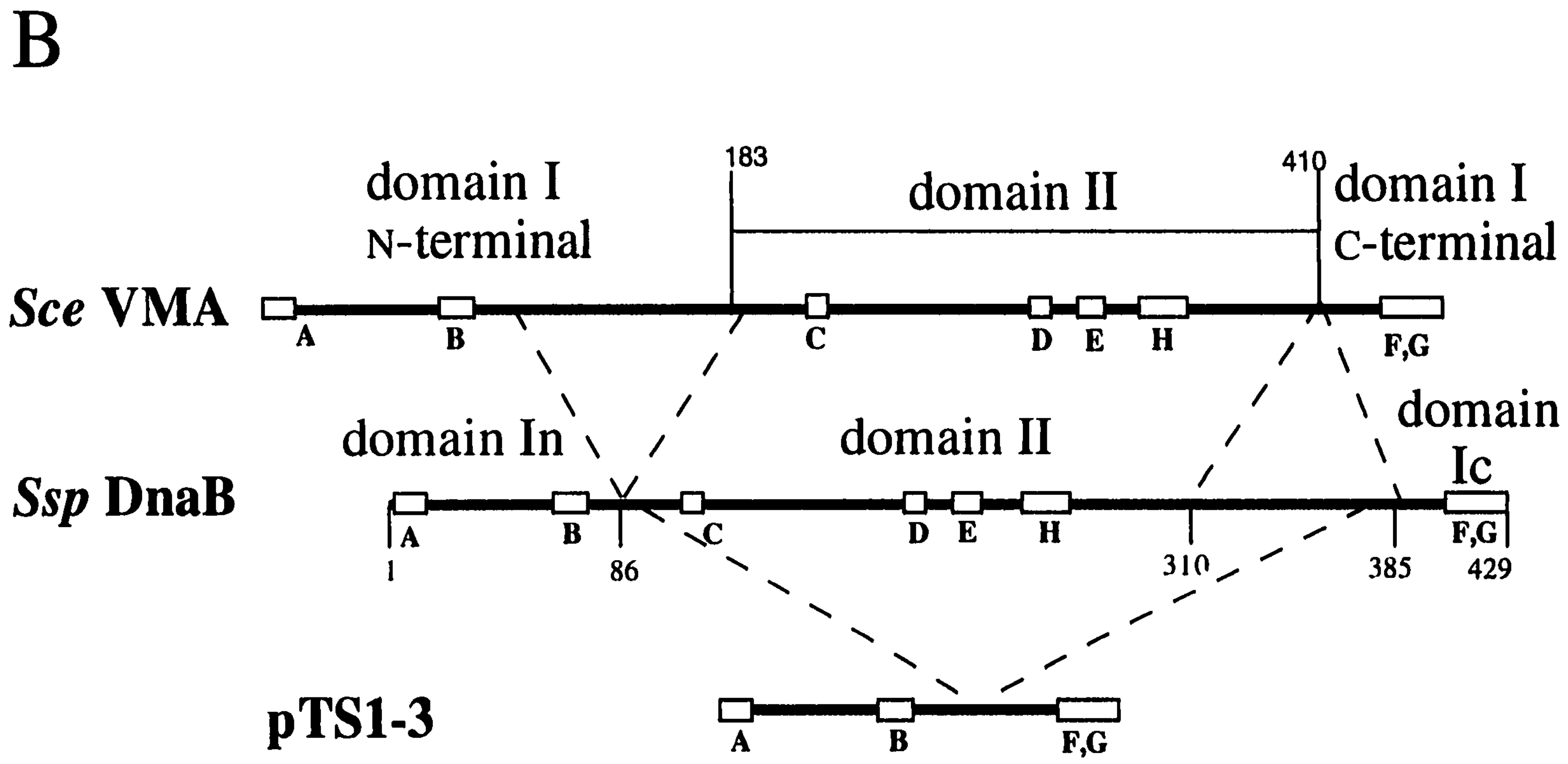

B
Sce VMA
Ssp DnaB
pTS1-3
domain I
N-terminal
183
domain II
410
domain I
C-terminal
A
B
C
D
E
H
F,G
domain In
domain II
domain
Ic
A
B
C
D
E
H
F,G
1
86
310
385
429
A
B
F,G

Figure 3-22

The size of the functional miniature intein is similar to the size of the *Porphyra purpurea* DnaB intein (*Ppu* DnaB intein, 150 aa in size). The *Ppu* DnaB intein, which contains motifs A, B, F, and G but completely lacks the dodecapeptide motifs C, E and motif D, could be a naturally occurring miniature intein, but it did not support protein splicing under conditions similar to those used for the *Ssp* DnaB intein (data not shown). Another natural mini-intein, the 198-aa GyrA intein in *Mycobacterium xenopi,* also lacks the endonuclease motifs. It has been shown to splice, but a linker sequence (in place of the endonuclease domain) as well as the native N-extein sequence are required for the splicing (Telenti *et al.*, 1997). Similar deletion mutants were also made for other inteins. In the *Mtu* RecA intein, the entire endonuclease domain could be deleted while retaining lower levels of splicing activity (Derbyshire *et al.*, 1997). In the *Sce* VMA intein, a portion of the endonuclease domain could be replaced with a linker polypeptide without abolishing the splice function, but deleting the entire endonuclease domain led to a loss of protein splicing activity (Chong and Xu, 1997). In the *Psp* Pol-1 intein, deletions of different sizes in the endonuclease domain all led to inactivation of splicing (Southworth *et al.*, 1998). The differences among these inteins suggest that the endonuclease domain and the native exteins of some (but not all) inteins may play a role in the correct folding and function of the splicing domain. In this respect, it is noted that the *Ssp* DnaB mini-inteins (pTS1-1 and pTS1-2) that have partial endonuclease domain sequences showed less efficient protein splicing in comparison to the mini-intein in pTS1-3 that lacks the entire endonuclease domain. The *Ssp* DnaB mini-intein flanked by native extein sequences (in pTS1-3) also showed less efficient protein splicing when compared to an identical mini-intein flanked by non-native exteins (in pMST). These observations suggest that the partial endonuclease domain and native extein sequences may actually interfere with the proper folding of the precursor protein or the splicing domain.

The functionally active mini-intein without the endonuclease domain also has implications regarding the evolution of inteins. Two models have been proposed on intein evolution. In both models, the invasive potential of homing endonuclease is likely to provide the basis of the maintenance and spread of inteins. According to one model, an endonuclease gene invaded a protein coding sequence and evolved protein splicing activity to keep the functional integrity of the host protein. In the second model, the endonuclease gene invaded the coding sequence of a protein splicing element (Duan *et al.*, 1997). The first model is supported by the observation that the HO endonuclease contains all of the intein motifs except the conserved splice junction residues (Pietrokovski, 1994). However, HO endonuclease is still unable to splice as an in-frame protein fusion after the addition of the junction residues (Perler *et al.*, 1997). Upon demonstration of the bipartite domain structure of the *Sce* VMA intein, Duan *et al.* hypothesized that inteins represent composite genes resulting from the invasion of an endonuclease open reading frame into a preexisting gene that encoded a protein splicing element (Duan *et al.*, 1997).

The functional mini-intein without its endonuclease domain as constructed here demonstrates a functional and structural independence between the protein-splicing domain and the endonuclease domains. This independence supports a separate origin of the two domains. Furthermore, the ability of the intein to function without the entire endonuclease domain favors the scenario in which the endonuclease gene invaded a preexisting intein-coding sequence. Otherwise some of the protein splicing function would be reside in the endonuclease domain. The symbiotic association of the endonuclease ORF with the splicing ORF benefits each other. The endonuclease ORF is associated with a gene encoding a protein that removes itself and the endonuclease from the host gene, preventing any effects on the host. The endonuclease, on the other hand, provides the means for inteins to be maintained and to spread among different genes, organisms, and possibly even kingdoms. This relationship is analogous to that in the proposed model that explains the relationship of group I introns and intron-encoded endonucleases (Belfort, 1989;

Lambowitz, 1989). That scenario suggests that the introns provide a means of excision of the endonuclease ORF by RNA splicing, and endonuclease ORF allows the intron to move to new locations in the genome. The endonuclease target site for the initial invasion has been found in a group I intron (Loizos *et al.*, 1994). The *sunY* intron-encoded I-TevII endonuclease of T4 phage can cleave a synthetic intron that lacks the endonuclease ORF. The fused intron-endonuclease could then move to new genomic locations. The double-strand-break-mediated gene conversion event could be the possible mode of entry of the endonuclease ORF. Nevertheless, no such recognition site has been observed in inteins. It would be premature to state with certainty that an endonuclease-free intein is the ancestral intein.

3.3.4 Reconstitution of a splicing domain through non-covalent interactions:

The observation that the large central region of the *Ssp* DnaB intein is not essential for protein splicing suggested the possibility of reconstituting an active protein splicing element from the two terminal sequences of the intein by non-covalent interactions. The N- and C-terminal portions of the *Ssp* DnaB mini-intein with respective exteins were produced as separate polypeptides by introducing in-frame translation termination and initiation codons into the intein coding sequence. Such a insertion split the mini-intein into two parts between motifs B and F, resulting in split intein fragments of 106 aa and 48 aa, respectively. The N-terminal portion and the C-terminal portion were in different reading frames, so that accidental translational read-through by nonsense suppression would not lead to a functional product.

Efficient protein *trans*-splicing of the split mini-intein construct pMST-split indicates that the N-terminal fragment (106 aa) and the C-terminal fragment (48 aa) of the *Ssp* DnaB intein can come together to form a functional splicing domain without assistance of the endonuclease domain or a covalent linkage between them. Crystal structures of both the *Sce* VMA intein and the *Mxe* GyrA intein reveal non-covalent interactions between the N- and C-terminal sequences of the inteins (Duan *et al.*, 1997; Klabunde *et al.*, 1998). The observation of protein *trans*-splicing with the split *Ssp* DnaB mini-intein suggests that non-covalent interactions between the N- and C-terminal sequences of this intein are sufficient to bring the two sequences into correct assembly or folding for the protein *trans*-splicing to occur. The N-extein (maltose-binding protein) and the C-extein (thioredoxin) are, by design, two separate and stable structural domains. They are not known to interact with each other and therefore are unlikely to contribute to the assembly of the two intein fragments. Consistent with this finding, protein *trans*-splicing was also observed after replacing the non-native exteins with native exteins of the *Ssp* DnaB intein (data not shown).

3.3.5 The internal residues of the *Ssp* DnaB mini-intein involved in protein splicing:

The mutagenesis study on the *Ssp* DnaB mini-intein revealed some important residues for protein splicing. Substitution of Cys1 by Ser blocked the splicing of the MS'T fusion protein but the production of the C-terminal cleavage products, indicating that this substitution blocks the N/S acyl shift at the N-terminal splice junction, which is consistent with the proposed catalytic role of Cys1 in the splicing reaction. Although Cys and Ser have similar structures, the different chemical properties of Cys and Ser may have different effects on protein splicing. The thiol group of Cys has a pKa of 8.3 while the hydroxyl group of Ser has a pKa of 13.7. At certain pH, the thiol group of Cys can be deprotonated to act as a nucleophile, but the hydroxyl group of Ser can not do this due to its high pKa. Therefore, the structurally conservative substitution has completely different effect on protein splicing.

Besides the junction residues, the internal residues were also investigated for their roles in splicing. Three single mutations were found to block protein splicing: G80D, I126T, and I150V. Among the three mutations, only I150V falls into one of the conserved sequence motifs (motif G). In the *Sce* VMA intein, the residue at this position belongs to a hydrophobic cluster which interacts with another hydrophobic cluster preceding the N-terminal junction (Nogami *et al.*, 1997). Such an interaction helps the proximal localization of the two splicing junctions, thus ensuring the progression of the splicing reaction. Although the I150V mutation maintains the hydrophobic feature of this cluster, the different sizes of the side chains may interrupt the interaction between the two clusters. The other two mutations are likely to affect the proper folding of the precursor protein. G80D changes Gly to the charged residue Asp. It may change the local structure of the folded precursor. I126T substitutes Ile with a residue with smaller but hydroxyl-containing side chain. It could also affect the local folding of the precursor protein.

3.3.6 The endonuclease activity of the *Ssp* DnaB intein:

The *Ssp* DnaB intein has a typical intein size of 429 aa and, like a typical intein, contains the two endonuclease motifs (motifs C and E). The two motifs are important features of homing endonucleases. They are spaced about 100 amino acids apart and are not responsible for the substrate specificity, since each endonuclease recognizes and cleaves at different sites. Instead, the acidic amino acids in the two motifs and another Lys are required to form the catalytic center that cleaves the DNA strands (Gimble and Stephens, 1995). In the *Ssp* DnaB intein, the acidic residues (Asp121 in motif C and Asp226 in motif E) are present at those conserved positions (Fig 3-23). The Lys residue (Lys201) that facilitates the formation of the catalytic triad is also found in motif D. However, no structural information is available for this intein, which may show a correct alignment of these residues. Endonuclease activity was not detected for this intein under the conditions used in this study. This absence of activity is possibly because of an accumulation of mutations in the intein. The crystal structure of the *Sce* VMA intein clearly shows that the intein can fold into two domains: the splicing domain and the endonuclease domain. The acidic residues are located inside the endonuclease domain, forming the catalytic core. However, the DNA-binding site of this enzyme appears to include a part of the splicing domain (Duan *et al.*, 1997). The *Ssp* DnaB intein is likely to have structure similar to that of the *Sce* VMA intein. Although it still contains the acidic residues in the two dodecapeptide motifs, the DNA-recognition and DNA-binding ability of this intein might have been abolished due to mutations in the regions responsible for such functions. On the other hand, the artificial substrate covers only 48 bp of the sequence flanking the intein insertion site. Although the intein-encoded endonucleases usually recognize a sequence of 12-40 bp, it is not impossible that the enzyme cleaves DNA at a site away from the intein insertion site. In fact, the group I introns in bacteriophage T4 do encode such endonucleases, although these enzymes lack the LAGLIDADG motifs (Michel and Dujon, 1986). Alternatively, although the conditions used here for the endonuclease activity assay have been used in the studies of other intein/intron related endonucleases, they might not be the optimum conditions for the *Ssp* DnaB intein. If the endonuclease activity of *Ssp* DnaB is very low, a more sensitive method might be required to detect it.

Similar results were also observed in the study of the *Ceu* ClpP intein. The chloroplast *clpP* gene of *Chlamydomonas eugametos* contains two large insertion sequences (Huang *et al.*, 1994). One of the two insertion sequences, IS2, has been shown to be an intein, although it has a Gly instead of the conserved His at the C terminus of the intein. To restore its protein splicing activity in *E. coli* cells required substitution of the penultimate Gly with His (Wang and Liu, 1997). This 456-aa intein contains the conserved intein motifs, including motifs C and E which are putative endonuclease motifs (Fig. 3-23). Endonuclease activity assays were carried out under conditions similar to those used for the *Ssp* DnaB intein. However, no endonuclease activity was detected under the conditions used for the assay (data not shown). The substitution of the highly conserved His at the C terminus of the intein by Gly suggests that the ClpP IS2 intein may

Figure 3-23 Sequence comparison of the conserved sequence motifs in the endonuclease domain. The sequences aligned here are from the inteins that are shown to be active endonucleases: the *Sce* VMA intein, *Psp* Pol1-1 intein, and *Tli* Pol-1 intein, *Tli* Pol-2 intein. These sequence motifs are aligned with the sequence motifs from the *Ssp* DnaB intein and the *Ceu* ClpP intein. The acidic residues in motifs C and E and the Lys in motif D are highlighted. They are suggested to be involved in the formation of the catalytic core.

Intein	Motif C		Motif D		Motif E		Motif H	
Sce VMA	LLGLWIG**D**G	219	V**K**NIPSFL	307	FLAGLIDS**D**G	327	TIHTSVRDGLVSLARSLGL	359
Psp Pol-1	LLGYYVS**E**G	289	N**K**RVPEVI	364	FLEGYFIG**D**G	384	TKSELLVNGLVLLLNSLGV	414
Tli Pol-1	LLGYYVS**E**G	290	N**K**RIPSVI	365	FLEAYFTG**D**G	385	TKSELLANQLVFLLNSLGI	415
Tli Pol-2	LVGLIVG**D**G	156	R**R**KIPEFM	234	FLRGLFSA**D**G	254	NIDADFLREVRKLLWIVGI	287
Ssp DnaB	LLGHLIG**D**G	122	E**K**FVPNQV	207	FLRHLWST**D**G	227	TSSEKLAKDVQSLLLKLGI	263
Ceu ClpP	FFGLWIA**N**G	151	N**K**YLPDWV	230	LLNSLCLG**N**C	250	STSERFANDVSRLALHAGT	281

Figure 3-23

be somewhat degenerate. It may have accumulated additional but less critical mutations in other parts of the intein. Although the IS2 intein retains many of the putative intein motifs, the degree of overall sequence similarity between IS2 and other inteins is low. In the two putative endonuclease motifs, the critical acidic residues are missing (Fig. 3-23). This may cause the loss of endonuclease activity. Again, the technical aspects of endonuclease activity assay discussed above should also be taken into consider.

3.3.7 Summary:

The work described above clearly shows that the *Synechocystis* sp. PCC6803 DnaB protein contains an active intein, which undergoes protein splicing either in the context of its native exteins or in non-native exteins. A functional 154-aa mini-intein was derived from this intein by deleting the centrally located 275 amino acids of the intein sequence. The fact that the mini-intein can splice leads to the proposal of a model that predicts a two-domain structure for this intein. The model suggests that the intein consists of two structural domains: a protein-splicing domain and an endonuclease domain. These two domains are functionally and structurally separable. The mini-intein was further split into two intein fragments, and efficient protein *trans*-splicing was observed. This finding indicates that the N- and C-terminal regions of the intein, whether physically linked or not, can come together to form the protein-splicing domain. The study also showed that beside the junction residues, some of the internal residues of the intein could be involved in the splicing activity. Although the *Ssp* DnaB intein still contains the endonuclease domain, no endonuclease activity was detected under the conditions studied.

Chapter IV Protein *trans*-splicing of a split intein in a split DnaE protein of *Synechocystis* sp strain PCC6803

4.1 Introduction:

Up to now, approximately 50 intein coding sequences have been found in over 20 different genes distributed among the nuclear and organellar genomes of eukaryotes, archaebacteria, and eubacteria. All these inteins have continuous sequences; most are 400-500 amino acids in size with a protein splicing domain and an endonuclease domain, while a few mini-inteins are about 150 amino acids in size with a splicing domain only. Three inteins were found previously in the cyanobacterium *Synechocystis* sp. strain PCC6803, including the *Ssp* DnaB intein in a DNA helicase (Pietrokovski, 1996; also see Chapter III of this thesis), the *Ssp* DnaX intein in the τ subunit of DNA polymerase III (Liu and Hu, 1997a), and the *Ssp* GyrB intein in a DNA gyrase B subunit (Dalgaard *et al.*, 1997a; Pietrokovski, 1998). In this study, a new intein (the *Ssp* DnaE intein) is found in *Synechocystis* sp. strain PCC6803. This intein is present in the DnaE protein. DnaE is the catalytic subunit of bacterial DNA polymerase III. In *E. coli*, the DNA polymerase III holoenzyme is the replicative polymerase responsible for the synthesis of the majority of the genome. DnaE (also known as α subunit), in addition to its catalytic role, also serves as an organization protein to hold the 18-protein holoenzyme complex together. Its N-terminal part contains the polymerase active site, while the C-terminal part interacts directly with the τ subunit to form a dimeric polymerase and with the β subunit that forms a sliding clamp on the DNA template (Kim and McHenry, 1996). This study shows that the *dnaE* gene in *Synechocystis* sp. PCC6803 is split into two separate genes by a break in an intein sequence. This *Ssp* DnaE intein, unlike previously reported inteins, not only lacks an endonuclease domain but also exists as a split intein encoded by two separate genes. This study further demonstrates that the products of the split *dnaE* gene can undergo protein *trans*-splicing to form an intact DnaE protein. Although several inteins have been artificially split and shown to splice *in trans* both *in vivo* and *in vitro* (Shingledecker *et al.*, 1998; Southworth *et al.*, 1998; Mills *et al.*, 1998; Wu *et al.*, 1998a; also see Chapter III of this thesis), a natural protein *trans*-splicing system provides a new perspective on this phenomenon. In an independent study, Gorbalenya also predicted this intein-containing split *dnaE* gene through sequence analysis (Gorbalenya, 1998).

4.2 Results:

4.2.1 Sequence analysis of the split *Synechocystis* sp. PCC6803 *dnaE* genes:

The complete genome sequence has been determined previously for *Synechocystis* sp. PCC6803 (Kaneko *et al.*, 1996), and a list of the gene content can be seen at the CyanoBase web site (http://www.kazusa.or.jp/cyano/cyano.html). In browsing through the CyanoBase, two separate open reading frames (ORFs slr0603 and sll1572) were noticed. The two open reading frames show significant sequence similarities to the *E. coli* DnaE protein (DNA polymerase III α subunit). Further analysis revealed that ORF slr0603 and ORF sll1572 are two members of a discontinuous *dnaE* gene. These ORFs have been subsequently named *dnaE-n* and *dnaE-c*, respectively (Fig. 4-1). The *dnaE-n* coding sequence is 2694 bp long and spans bases 3,561,946 to 3,564,639 of the genome. The *dnaE-c* coding sequence is 1377 bp long and spans bases 737,811 to 736,435 of the genome. These two genes are separated by 745,226 bp of sequence and numerous unrelated genes on the 3,573,470-bp circular genome. In addition to distance, the coding sequences of these two genes are located on opposite DNA strands. There is no indication of intronic sequence either downstream of *dnaE-n* or upstream of *dnaE-c*. In fact, the *dnaE-n* gene is followed immediately downstream by a *lepA* gene that encodes a GTP-binding protein unrelated to the DnaE protein, with a 199-bp intergenic spacer between them. The *dnaE-c* gene is flanked upstream by an unidentified open reading frame that is unrelated to the DnaE protein and has some similarity to lysostaphin, with a 215-bp intergenic spacer between them. There is no additional DnaE-like gene listed in the CyanoBase. Extensive blast searches of the complete *Ssp* genome sequence did not show any additional *dnaE* gene (complete or in fragments).

Protein sequence deduced from the *dnaE-n* gene can be divided into two regions: a 774-aa extein region named Ext-n followed by a 123-aa intein region named Int-n.

Figure 4-1 Gene map of the *Ssp* dnaE genes. Two members of the split *dnaE* gene, *dnaE-n* and *dnaE-c*, are shown on the genome of *Synechocystis* sp. PCC6803 (*Ssp* genome).

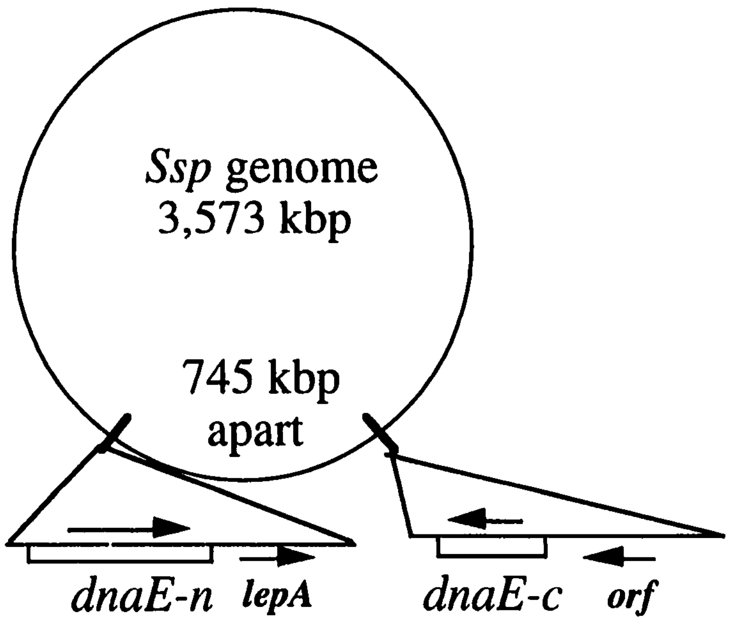


Figure 4-1

Figure 4-2 Protein structure of the *Ssp* DnaE proteins. In the predicted proteins, DnaE-related sequences (dotted boxes) are specified as exteins Ext-n and Ext-c, while intein-related sequences (black boxes) are specified as Int-n and Int-c. The exteins are related to *E. coli* DnaE protein, whose functional domains are marked.

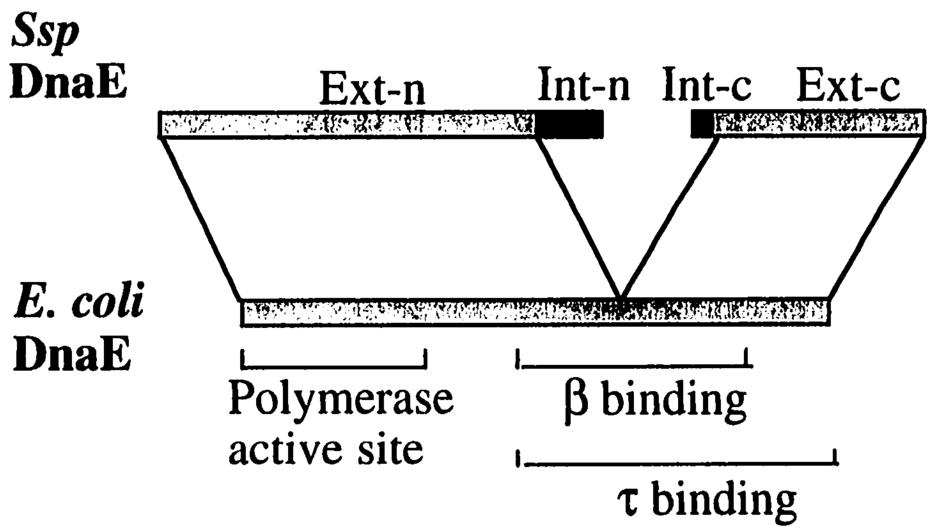

Ssp
DnaE
Ext-n
Int-n
Int-c
Ext-c
E. coli
DnaE
Polymerase
active site
β binding
τ binding

Figure 4-2

Similarly, protein sequences deduced from the *dnaE-c* gene can be divided into a 36 aa intein region (Int-c), followed by a 423-aa extein region (Ext-c). The Ext-n and Ext-c sequences correspond to the N- and C-terminal halves of a DnaE protein, respectively, and together they reconstitute a complete DnaE sequence (Fig. 4-2). This *Ssp* DnaE sequence, although discontinuous, resembles the continuous DnaE sequences of other organisms both in length and in sequence (Fig. 4-3). The *Ssp* DnaE sequence is 36%, 37%, and 35% identical to DnaE proteins of *E. coli, Bacillus subtilis, and Mycobacterium tuberculosis*, respectively, over the entire 1,196-aa sequence. These degrees of sequence identity are comparable to the 35-36% sequence identities found among DnaE proteins of the other three compared bacterial organisms.

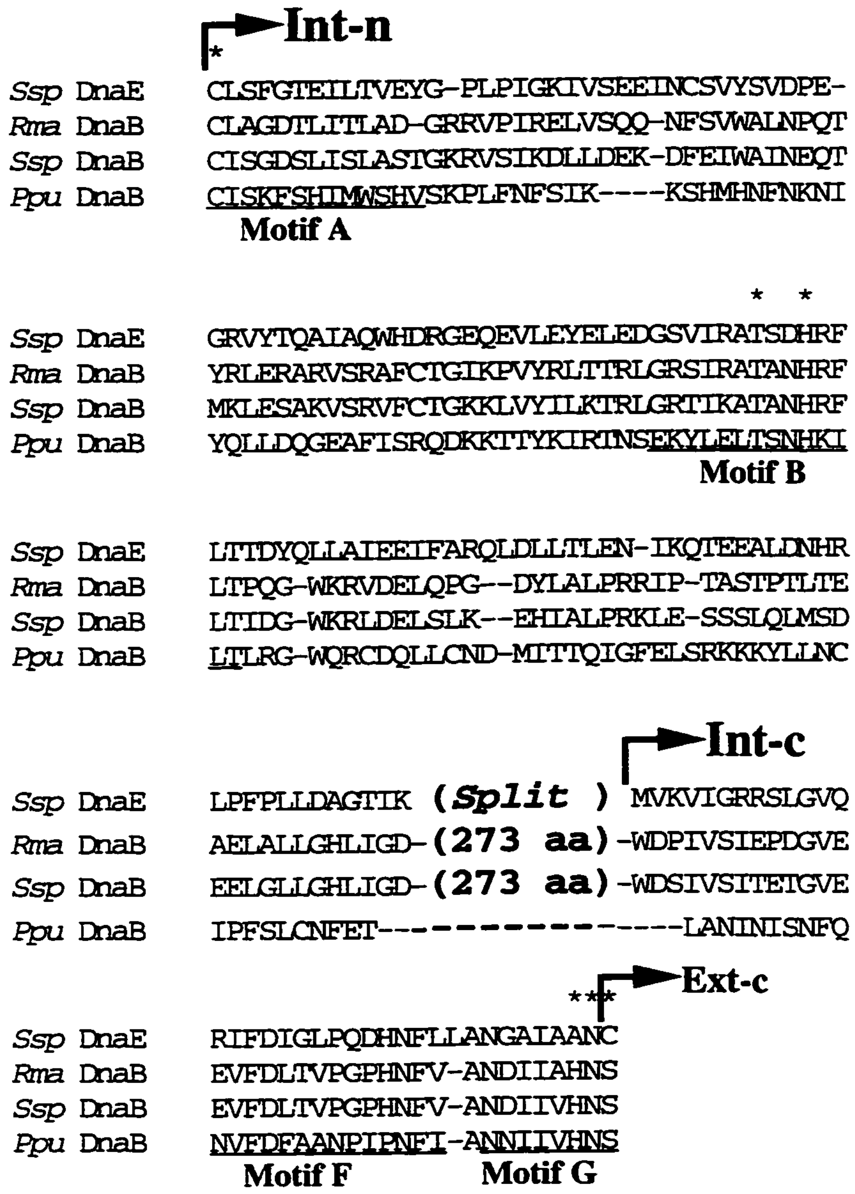
The Int-n and Int-c sequences show no detectable similarity to DnaE proteins, but instead have marked similarity to known intein sequences (Fig. 4-4). Int-n and Int-c correspond to the N- and C-terminal halves of the intein, and together they reconstitute a mini-intein sequence (named the *Ssp* DnaE intein) with a composite length of 159 aa. The sequence of this discontinuous *Ssp* DnaE intein is most similar to corresponding sequences of the *Rma* DnaB intein found previously in a DnaB protein (DNA helicase) of the thermophilic eubacterium *Rhodothermus marinus* (Liu and Hu, 1997b). The *Ssp* DnaE intein sequence is 30% identical to the *Rma* DnaB intein and 22% identical to the *Ssp* DnaB intein over the 159-aa sequence. Much lower sequence identities were found in comparison with other known inteins. The *Ssp* DnaE intein, in addition to being split, lacks sequences for a centrally located endonuclease domain that is present in most known inteins, including the *Rma* DnaB intein. Nevertheless, the split *Ssp* DnaE intein has many known sequence features of an intein splicing domain. A 50% sequence identity was found between the *Ssp* DnaE intein and the *Rma* DnaB intein over the conserved sequence blocks (A, B, F, and G, totaling 49 aa). Residues important for the catalysis of protein splicing were found in the *Ssp* DnaE intein, including a nucleophilic

Figure 4-3 Sequence comparison of DnaE proteins. The *Ssp* DnaE extein sequences (*Ssp*) are aligned with corresponding DnaE sequences of *E. coli* (*Eco*), *Bacillus subtilis* (*Bsu*), and *Mycobacterium tuberculosis* (*Mtu*). Only sequences proximal to the intein sequences (Int-n and Int-c) are shown, while the number of omitted residues at the N- and C-termini are shown in parentheses. Symbols: - represent gaps introduced to optimize the alignment; * and . mark positions of identical and similar amino acids, respectively.

```
Ssp   (633 aa)-FQLESQGMKQIVRDLKPSGIEDISSILALYRPGPLDAG
Eco   (608 aa)-FQLESRGMKDLIKRLQPDCFEDMIALVALFRPGPLQSG
Bsu   (584 aa)-FQLESAGMRSVLKRLKPSGLEDIVAVNALYRPGPMEN-
Mtu   (636 aa)-FQLDGGPMRDLLRRMQPTGFEDVVAVIALYRPGPMGMN
               ***.   *. ... ..*   **. .. **.****.

Ssp   LIPIFINRKHGRE-----EISYQHKLLEPILNETYGVLVYQEQIMKM
Eco   MVDNFIDRKHGREEISYPDVQWQHESLKPVLEPTYGIILYQEQVMQI
Bsu   -IPLFIDRKHGRA-----PVHYPHEDLRSILEDTYGVIVYQEQIMMI
Mtu   AHNDYADRKNNRQ-AIKPIHPELEEPLREILAETYGLIVYQEQIMRI
          .  **..*              *  .*  ***...****.* .

Ssp   AQDLADYSLGEADLLRRAMGKKKAEEMQKHRAKFVDGSTKHGVPSRI
Eco   AQVLSGYTLGGADMLRRAMGKKKPEEMAKQRSVFAEGAEKNGINAEL
Bsu   ASRMAGFSLGEADLLRRAVSKKKKEILDRERSHFVEGCLKKEYSVDT
Mtu   AQKVASYSLARADILRKAMGKKKREVLEKEFEGFSDGMQANGFSPAA
      *  .. ..*  **.**.*. *** * . .    * .*

Ssp   AENLFDQMVKFAEY [Int-n]  [Int-c] CFNKSHSTAYA
Eco   AMKIFDLVEKFAGY----------------------GFNKSHSAAYA
Bsu   ANEVYDLIVKFANY----------------------GFNRSHAVAYS
Mtu   IKALWDTILPFADY----------------------AFNKSHAAGYG
         . * .  ** *                       **.**.  *

Ssp   YVTYQTAYLKANYPVEYMAALLTASSDSQEKVEKYRENCQKMGITVE
Eco   LVSYQTLWLKAHYPAEFMAAVMTADMDNTEKVVGLVDECWRMGLKIL
Bsu   MIGCQLAYLKAHYPLYFMCGLLTSVIGNEDKISQYLYEAKGSGIRIL
Mtu   MVSYWTAYLKANYPAEYMAGLLTSVGDDKDKAAVYLADCRKLGITVL
       .      ***.**  .*. ..*.     .*       .   *. .

Ssp   PPDINRSQRHFTPLG-EAILFGLSAVRNLGEGAIEQIITARDNSEEK
Eco   PPDINSGLYHFHVNDDGEIVYGIGAIKGVGEGPIEAIIEARN--KGG
Bsu   PPSVNKSSFPFTVEN-GSVRYSLRAIKSVGVSAVKDIYKAR---KEK
Mtu   PPDVNESGLNFASVG-QDIRYGLGAVRNVGANVVGSLLQTRN--DKG
      ** .*     *       . . . *.. .*   .  .  .*

Ssp   RFKSLADFCTQVDLRVVNRRAIETLIMAGAFD-(286 aa)
Eco   YFRELFDLCARTDTKKLNRRVLEKLIMSGAFD-(271 aa)
Bsu   PFEDLFDFCFRVPSKSVNRKMLEALIFSGAMD-(201 aa)
Mtu   KFTDFSDYLNKIDISACNKKVTESLIKAGAFD-(269 aa)
       *    *   .      *..  * **  .** *
```

Figure 4-3

Figure 4-4 Sequence comparison of inteins. The *Ssp* DnaE intein sequences (*Ssp* DnaE), consisting of Int-n and Int-c as indicated, are aligned with corresponding sequences of *Rhodothermus marinus* DnaB intein (*Rma* DnaB), *Synechocystis* sp. PCC6803 DnaB intein (*Ssp* DnaB), and *Porphyra purpurea* chloroplast DnaB intein (*Ppu* DnaB). In the *Rma DnaB* intein and the *Ssp* DnaB intein, only sequences relating to Int-n and Int-c are shown, while the number of omitted residues are shown in parentheses. Putative intein motifs (Blocks A, B, F, and G) are underlined, with several critical residues marked by *.



```
               Int-n
               *
Ssp DnaE       CLSFGTEILTVEYG-PLPIGKIVSEEINCSVYSVDPE-
Rma DnaB       CLAGDTLITLAD-GRRVPIRELVSQQ-NFSVWALNPQT
Ssp DnaB       CISGDSLISLASTGKRVSIKDLLDEK-DFEIWAINEQT
Ppu DnaB       CISKFSHIMWSHVSKPLFNFSIK----KSHMHNFNKNI
                 Motif A

                                                 *  *
Ssp DnaE       GRVYTQAIAQWHDRGEQEVLEYELEDGSVIRATSDHRF
Rma DnaB       YRLERARVSRAFCTGIKPVYRLTTRLGRSIRATANHRF
Ssp DnaB       MKLESAKVSRVFCTGKKLVYILKTRLGRTIKATANHRF
Ppu DnaB       YQLLDQGEAFISRQDKKTTYKIRINSEKYLELTSNHKI
                                             Motif B

Ssp DnaE       LTTDYQLLAIEEIFARQLDLLTLEN-IKQTEEALDNHR
Rma DnaB       LTPQG-WKRVDELQPG--DYLALPRRIP-TASTPTLTE
Ssp DnaB       LTIDG-WKRLDELSLK--EHIALPRKLE-SSSLQLMSD
Ppu DnaB       LTLRG-WQRCDQLLCND-MITTQIGFELSRKKKYLLNC

                                          Int-c
Ssp DnaE       LPFPLLDAGTIK (Split ) MVKVIGRRSLGVQ
Rma DnaB       AELALLGHLIGD-(273 aa)-WDPIVSIEPDGVE
Ssp DnaB       EELGLLGHLIGD-(273 aa)-WDSIVSITETGVE
Ppu DnaB       IPFSLCNFET-----------------LANINISNFQ

                                 ***   Ext-c
Ssp DnaE       RIFDIGLPQDHNFLLANGAIAANC
Rma DnaB       EVFDLTVPGPHNFV-ANDIIAHNS
Ssp DnaB       EVFDLTVPGPHNFV-ANDIIVHNS
Ppu DnaB       NVFDFAANPIPNFI-ANNIIVHNS
                 Motif F        Motif G
```

Figure 4-4

residue (Cys) at the beginning of the intein sequence, another Cys at the beginning of C-extein, a Thr and a His in sequence motif B, and an Asn at the end of the intein. An Ala precedes the C-terminal Asn in the *Ssp* DnaE intein, while this position is occupied by His in most, but not all, other known inteins.

The insertion site of the split *Ssp* DnaE intein is inside the β– and τ– binding domains but outside the polymerase active site of the DnaE protein, according to a comparison with the better studied *E. coli* DnaE protein (Fig. 4-2). The *Ssp* DnaE intein disrupts a conserved region of the DnaE sequence (Fig. 4-3), which helped to define the extein-intein boundaries. The first residue of Ext-c in the *Ssp* DnaE sequence is Cys, while this position in the other DnaE proteins is occupied by Gly or Ala. This substitution is consistent with a requirement of the Cys in *Ssp* DnaE for protein splicing and the absence of an intein in the other DnaE proteins.

4.2.2 Overproduction of DnaE-n and DnaE-c proteins in *E. coli*, and preparation of antibodies:

The *Ssp dnaE-n* and *dnaE-c* genes were amplified by PCR from *Synechocystis* sp. PCC6803 total DNA as described in section 2.5.1. The two genes were separately inserted into expression vector pET-32, placing each gene behind a T7 promoter and the coding sequences of the *E. coli* thioredoxin gene, poly-His tag and S-tag. The resulting recombinant plasmids (pHC-199 and pHC-201, containing *dnaE-n* and *dnaE-c* genes, respectively) were transformed into *E. coli* strain BL21(DE3) cells, which harbor an IPTG-inducible T7 RNA polymerase gene. The *E. coli* cells were grown in liquid medium and induced to produce the fusion proteins by the addition of IPTG. Overproduction of the DnaE-n and DnaE-c fusion proteins was observed after three-hour induction (Fig. 4-5). pHC-199 produced a fusion protein containing the full-length DnaE-n protein with a size of 122 kDa. pHC-201 produced a fusion protein (57 kDa)

Figure 4-5 Overproduction and purification of DnaE-n and DnaE-c proteins. Upper: schematic illustration of fusion proteins encoded in the corresponding plasmids. pHC-199 encodes the full-length DnaE-n protein, while pHC-201 encodes the partial DnaE-c protein. In each case, the DnaE intein (hatched box) and extein (solid box) are fused in-frame with vector-encoded sequences (open box). T, H, and S stand for thioredoxin, poly-histidine tag, and S-tag, respectively. Lower: Protein gels. Total protein of uninduced cells (lanes 1 and 4), induced cells (lanes 2 and 5), and TALON resin-purified proteins (lanes 3 and 6) were resolved by SDS-polyacrylamide gel electrophoresis and visualized by Coomassie Blue staining. The discrepancy on the migration of the proteins in lanes 1 and 4 are possibly due to the fact that the proteins were produced in different inductions and run on different gels.

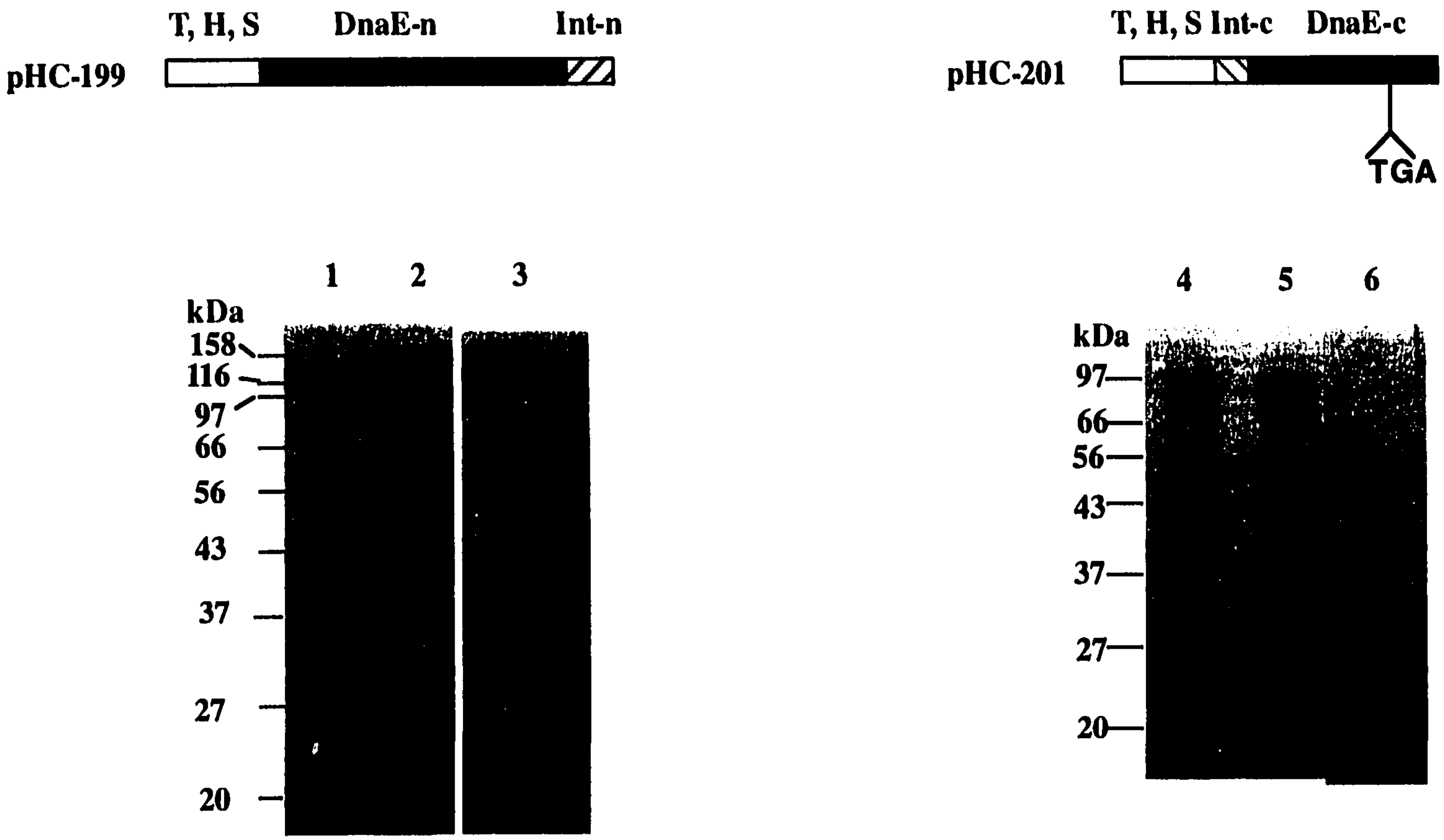

T, H, S
DnaE-n
Int-n
pHC-199
T, H, S Int-c
DnaE-c
pHC-201
TGA
1
2
3
kDa
158
116
97
66
56
43
37
27
20
4
5
6
kDa
97
66
56
43
37
27
20

Figure 4-5

containing a truncated DnaE-c protein, due to a PCR-introduced mutation 129 aa from the C terminus of the protein. Both proteins were found in the insoluble fraction of the cell lysate. After solublization by urea, the overproduced fusion proteins were isolated by using TALON metal affinity resin as described in section 2.9.1.c (Fig. 4-5), taking advantage of the fact that each fusion contains a poly-histidine sequence that binds the metal affinity resin. This purification step resulted in nearly pure DnaE-n and DnaE-c proteins. To remove minor contaminating *E. coli* proteins, the isolated proteins were resolved on preparative SDS-polyacrylamide gels and stained with Coomassie Blue. The gel slices containing the fusion proteins were cut out and used in antibody production. Specificities of the resulting antisera were confirmed by testing on the corresponding antigen.

4.2.3 Demonstration of protein *trans*-splicing of the *Ssp* DnaE intein in *E. coli*:

a. Construction of a *Ssp dnaE* operon for study of splicing activity:

As described in Chapter III, an intein with continuous sequence can be engineered to a functional split intein. The *Ssp* DnaE intein, unlike the other known inteins, has a discontinuous sequence, although it has all the critical sequence motifs forming the splicing domain. Whether this split intein is still able to support protein splicing becomes an interesting question. To test its protein *trans*-splicing activity in *E. coli* cells, the *Ssp* DnaE-n and DnaE-c coding sequences were inserted into the expression plasmid vector pET-32 to form a two-gene operon, allowing production of the two proteins inside the same *E. coli* cell (Fig. 4-6). The construct, pHC-209, was achieved by the PCR-mediated method described in section 2.5.3 using a pair of oligonucleotide primers: 5'-TTAATAA-TAATGGGTACCTTGAAAATGGATTTTTTAGGCTTG and 5'-ATTATTATTAA-CCTCCTTAACTCTGGCTTTGGGGTAACAGTGG. pHC-209 contains the complete DnaE-c coding sequence (1377 bp) and a partial DnaE-n coding sequence (named dnaE-n', 1017 bp). A complete DnaE-n coding sequence was not used, because it resulted in

Figure 4-6 Schematic illustration of a two-gene operon for co-expression of *dnaE-n* and *dnaE-c* genes in *E. coli* cells. The genes are constructed as a two-gene operon in an expression plasmid vector, with the complete DnaE-c coding sequence followed by a partial DnaE-n coding sequence (DnaE-n'). In the intergenic spacer, the termination codon (*TAA*) of DnaE-c and initiation codon of DnaE-n' are boxed, and the Shine-Dalgarno sequence (ribosome-binding site) is underlined. Products of the two genes are shown as precursor proteins, with their extein regions (Ext-n' and Ext-c) and intein regions (Int-n and Int-c) as indicated. Protein *trans*-splicing produces a spliced protein and excised intein fragments.

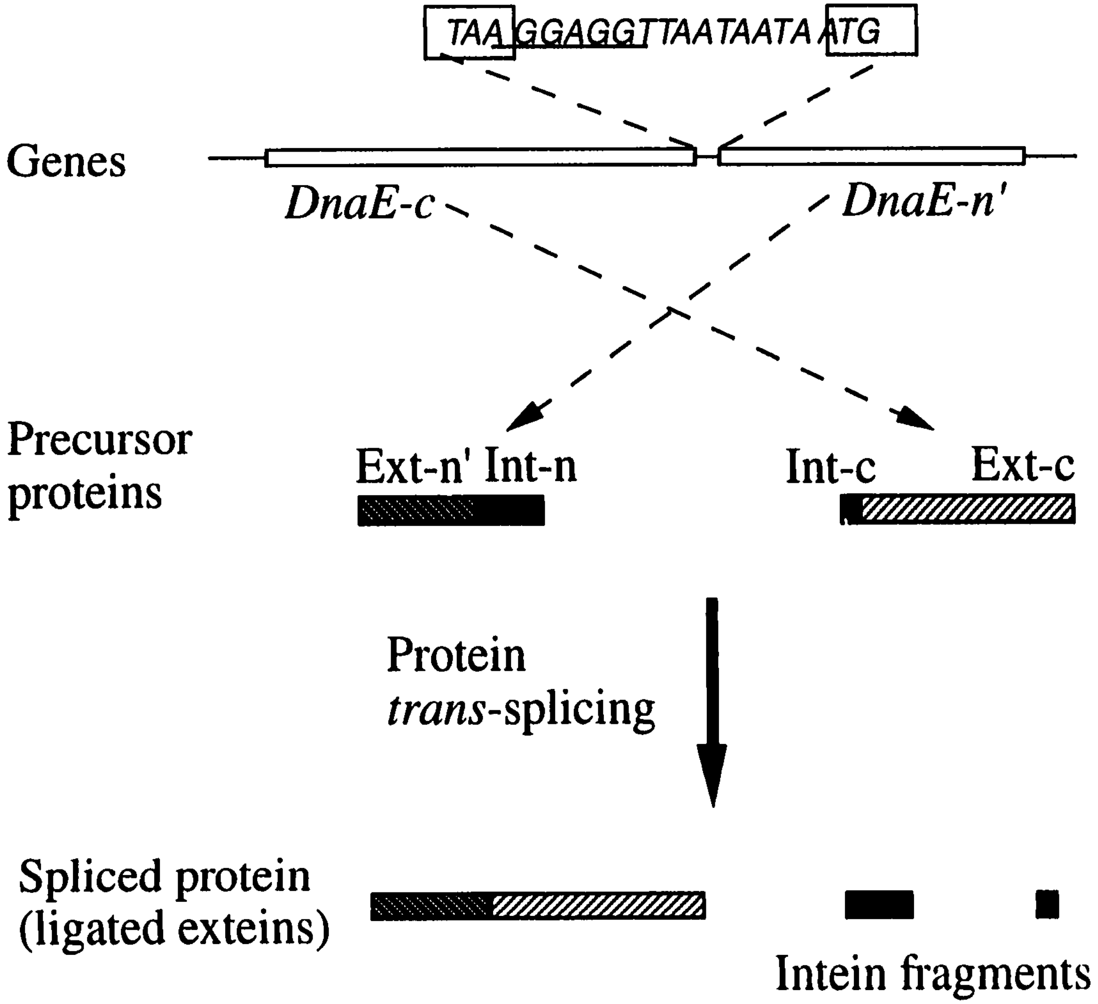

TAA GGAGGTTAATAATA ATG
Genes
DnaE-c
DnaE-n'
Precursor proteins
Ext-n' Int-n
Int-c
Ext-c
Protein trans-splicing
Spliced protein (ligated exteins)
Intein fragments

Figure 4-6

lower production and elevated degradation (fragmentation) of the protein product (data not shown). The partial DnaE-n sequence, DnaE-n', consists of a portion of the Ext-n sequence (216 aa, proximal to the intein) followed by the entire Int-n sequence. The DnaE-c and DnaE-n' coding sequences are separated by a small intergenic spacer which contains a cassette of termination codon - Shine-Dalgarno sequence - initiation codon. The sequence of the cassette was confirmed by DNA sequencing. In this construct, the DnaE-c and the DnaE-n' coding sequence are in different reading frames, and the DnaE-c coding sequence was placed in front of the DnaE-n' coding sequence to prevent accidental fusion of the split intein sequences which might arise through accidental translation of the small intergenic spacer.

b. Characterization of the *Ssp* DnaE intein for its protein *trans*-splicing activity in *E. coli*:

E. coli cells containing the above recombinant plasmid were induced by adding IPTG to produce the DnaE-c protein, the DnaE-n' protein, and possibly a spliced protein. After induction, total cellular proteins were resolved by analytical SDS-polyacrylamide gel electrophoresis and stained with Coomassie Blue. The results are shown in Fig. 4-7. For plasmid pHC-209, three protein products (C, N, and N-C) were observed after induction (Fig. 4-7, lane 2). Protein C and protein N were identified as the precursor proteins DnaE-c and DnaE-n', respectively. Their apparent sizes matched closely the predicted sizes of the precursor proteins (51 kDa for protein C and 38 kDa for protein N). In Western blot analysis, the two protein bands were specifically recognized by antiserum raised against the corresponding protein (Fig. 4-7, lane 4, 6). The third protein, N-C, has an apparent size of 71 kDa, which is consistent with the predicted size of the spliced protein. This protein was further analyzed by Western blot analysis. It was specifically recognized by antibodies raised against both the DnaE-n protein and the DnaE-c protein. This indicated that protein N-C contains both DnaE-n and DnaE-c protein sequences. This protein is very likely to be the spliced protein (ligated exteins).

Figure 4-7 Protein *trans*-splicing of the *Ssp* DnaE proteins. Total proteins of uninduced cells (lanes 1, 3, and 5) and induced cells (lanes 2, 4, and 6) were resolved by SDS-polyacrylamide gel electrophoresis and visualized by Coomassie Blue staining (lane 1 and 2), by Western blotting with anti-N (DnaE-n) antiserum (lane 3 and 4), or by Western blotting with anti-C (DnaE-c) antiserum (lane 5 and 6). The protein samples from lanes 1 and 2 were diluted 50 fold before using in the Western blot analysis. Positions of precursor proteins (N and C) and the spliced protein (N-C) are marked.

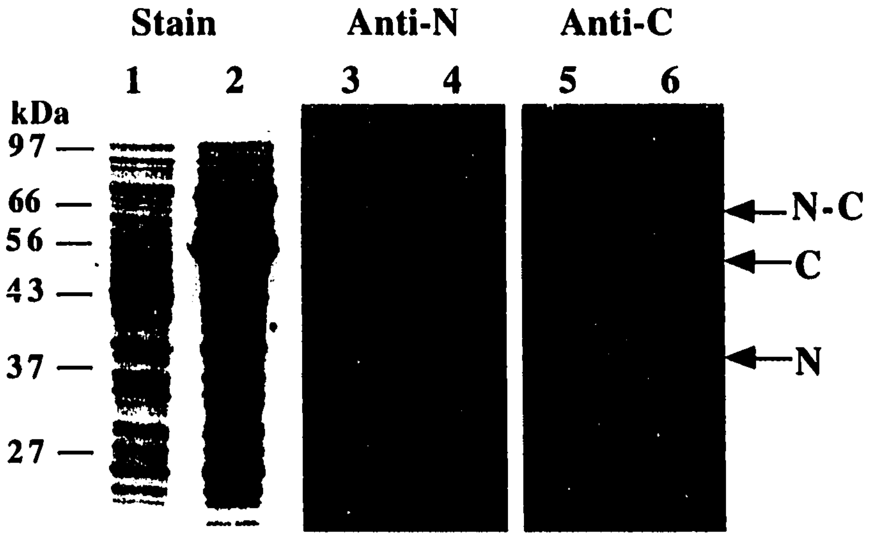


Figure 4-7

c. Identification of the 71-kDa protein as the spliced protein by protein micro-sequencing:

To further investigate the identity of the 71-kDa protein produced from the recombinant plasmid pHC-209, this protein was purified from the *E. coli* cells. Although this protein is soluble, it does not have an affinity tag that could be used for affinity purification. As an alternative, excision of a gel slice containing the 71-kDa protein band followed by electroelution of the protein from the gel proved to be an effective way of purifying this protein. The contamination from the co-migrated *E. coli* protein can be reduced to minimum by tight cutting of the gel. The purification results are shown in Fig. 4-8.

A sufficient amount of the 71-kDa protein was purified in order to carry out protein micro-sequencing. The purified 71-kDa protein was electrophoresed on an SDS-polyacrylamide gel, blotted onto a PVDF membrane, and stained briefly with Ponceau S to visualize the protein band. The 71-kDa protein band was excised from the membrane and used in protein micro-sequencing.

The 71-kDa protein was first subjected to N-terminal sequencing, revealing a 17-aa sequence, KMDFLGLKNLTTLQRAV, which matched precisely the predicted DnaE-n' sequence at amino acid positions 5 to 21 (Fig. 4-11). Amino acids at positions 2 to 4 were not determined due to sequencing failures at these positions, and the N-terminal f-Met apparently had been removed in the *E. coli* cell. The 71-kDa protein was further treated with the protease trypsin to cleave it into small pieces suitable for protein sequencing. This step generated the small fragment that spans the predicted splice junction. After trypsin treatment, the resulting peptide fragments were resolved by high performance liquid chromatography (HPLC) (Fig. 4-9). Three fragments (#83, #107, #112) were selected as targets based on their elution positions and UV absorption spectra. These fragments were then subjected to mass spectrometric analysis to determine their

Figure 4-8 Purification of the putative spliced protein by electroelution. All samples were resolved on SDS-polyacrylamide gel and stained with Coomassie Blue. Lane 1, total cellular protein of uninduced cells. Lane 2, total cellular protein of induced cells. Lane 3, purified putative spliced protein.

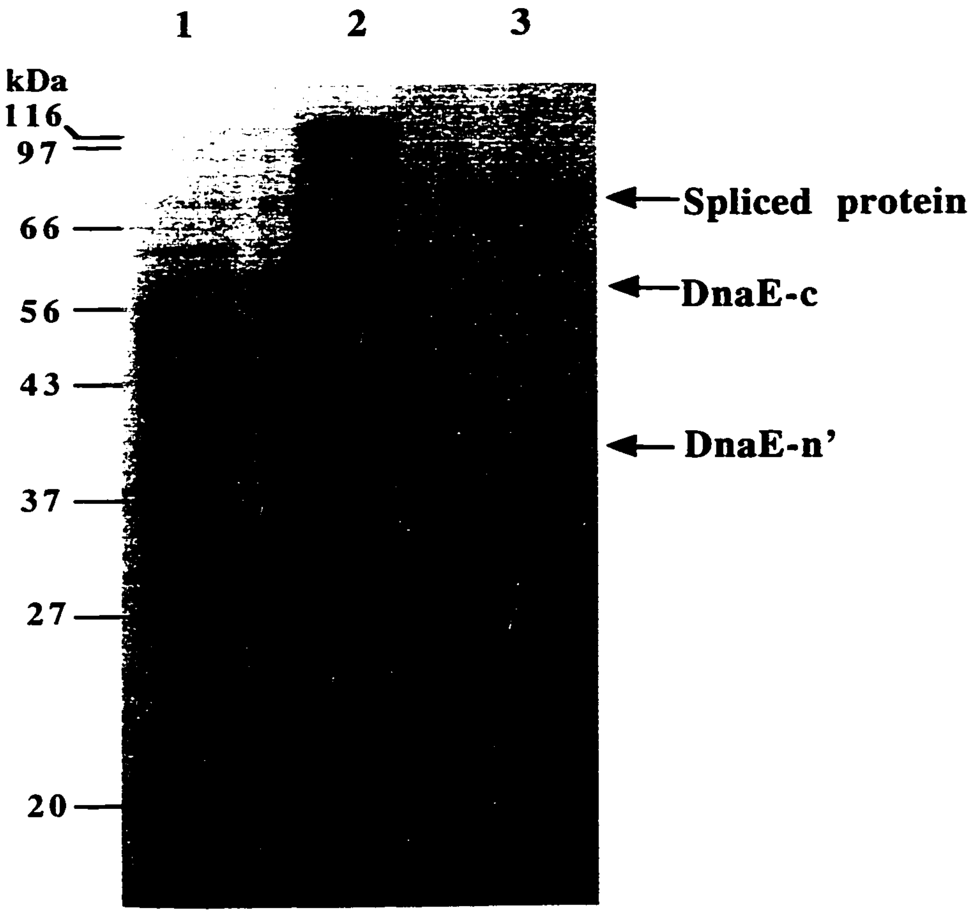


Figure 4-8

Figure 4-9 Isolation of tryptic polypeptides by HPLC. The putative spliced protein was digested with trypsin. The resulting polypeptides were resolved by a C-18 HPLC column. Top: absorbence of each tryptic fragment at 205 nm (the absorbence of amide). The height of each peak represent the size of each fragment. Bottom: Absorbence of each fragment at 277 nm or 292 nm (the absorbence of aromatic residues). Three polypeptides possiblely spanning the splice junction were selected in the analysis based on their elution positions and UV absorption spectra. The selected polypeptides (#83, #107, #112) are marked by arrows.

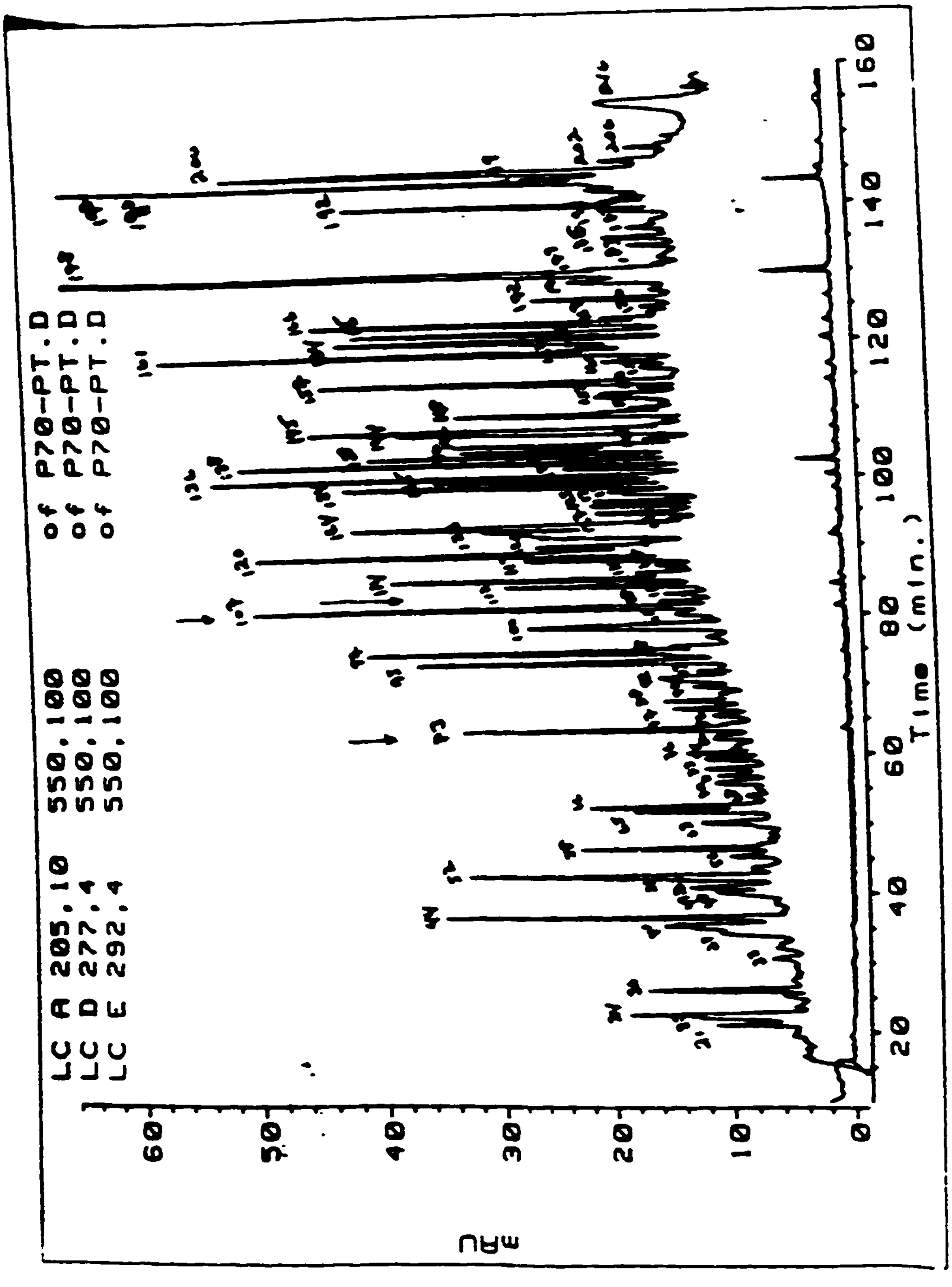

LC A 205,10 550,100 of P70-PT.D
LC D 277,4 550,100 of P70-PT.D
LC E 292,4 550,100 of P70-PT.D
mAU
0
10
20
30
40
50
60
20
40
60
80
100
120
140
160
Time (min.)

Figure 4-9

precise masses (Fig. 4-10). Two polypeptides (#112 and #107, corresponding to peptides III and IV in Fig. 4-11) inside the DnaE-c sequence were identified by matching their molecular masses to predicted molecular masses: for peptide III, the predicted mass is 1967.2 while the measured mass is 1965.0; for peptide IV, the predicted mass is 1555.8 while the measured mass is 1556.4. Peptide III corresponded to the sequence SHSTAYAYVTYQTAYLK (amino acid positions 220 to 236), while peptide IV corresponded to the sequence EHLGFYVSEHPLK (amino acid positions 428 to 440). Most importantly, a polypeptide (#83, corresponding to peptide II in Fig. 4-11) spanning the spliced junction was identified and sequenced. Its sequence, FAEYCFNK, matches precisely the predicted sequence in a spliced protein, with the sequence FAEY being the last four residues of Ext-n' and the sequence CFNK being the first four residues of Ext-c. This result indicates precise excision of the intein sequences (Int-n and Int-c) and joining of the extein sequences (Ext-n' and Ext-c) by a normal peptide bond. The two excised intein fragments were predicted but not observed, most likely due to their small sizes (14 kDa for Int-n and 4 kDa for Int-c), weak binding by the anti-N and anti-C antisera, and / or rapid degradation in *E. coli* cells. Nevertheless, production of the spliced protein (protein N-C) demonstrates that protein *trans*-splicing had occurred. Comparing the amount of protein N-C and the amount of protein N indicates that approximately 80% of the precursor protein N was incorporated into the spliced protein. The remaining protein N may have misfolded. Protein C was accumulated much more than protein N, indicating that the *dnaE-c* gene was expressed much more than the downstream *dnaE-n'* gene, probably as a result of inefficient translational coupling of the two-gene operon or a more rapid degradation of protein N.

4.2.4 Study of protein *trans*-splicing of the DnaE intein in *Synechocystis* sp. PCC6803:

In the CyanoBase web site, the split *Ssp dnaE* gene is the only *dnaE*-like gene listed. Extensive Blast searches of the complete *Ssp* genome sequence did not find any

Figure 4-10 Mass determination of selected polypeptides. Mass spectrometry of the selected peptides (#83, #107, #112) are shown. **A.** Mass spectrometry of peptide #83. **B.** Mass spectrometry of peptide #107. **C.** Mass spectrometry of peptide #112.

A.

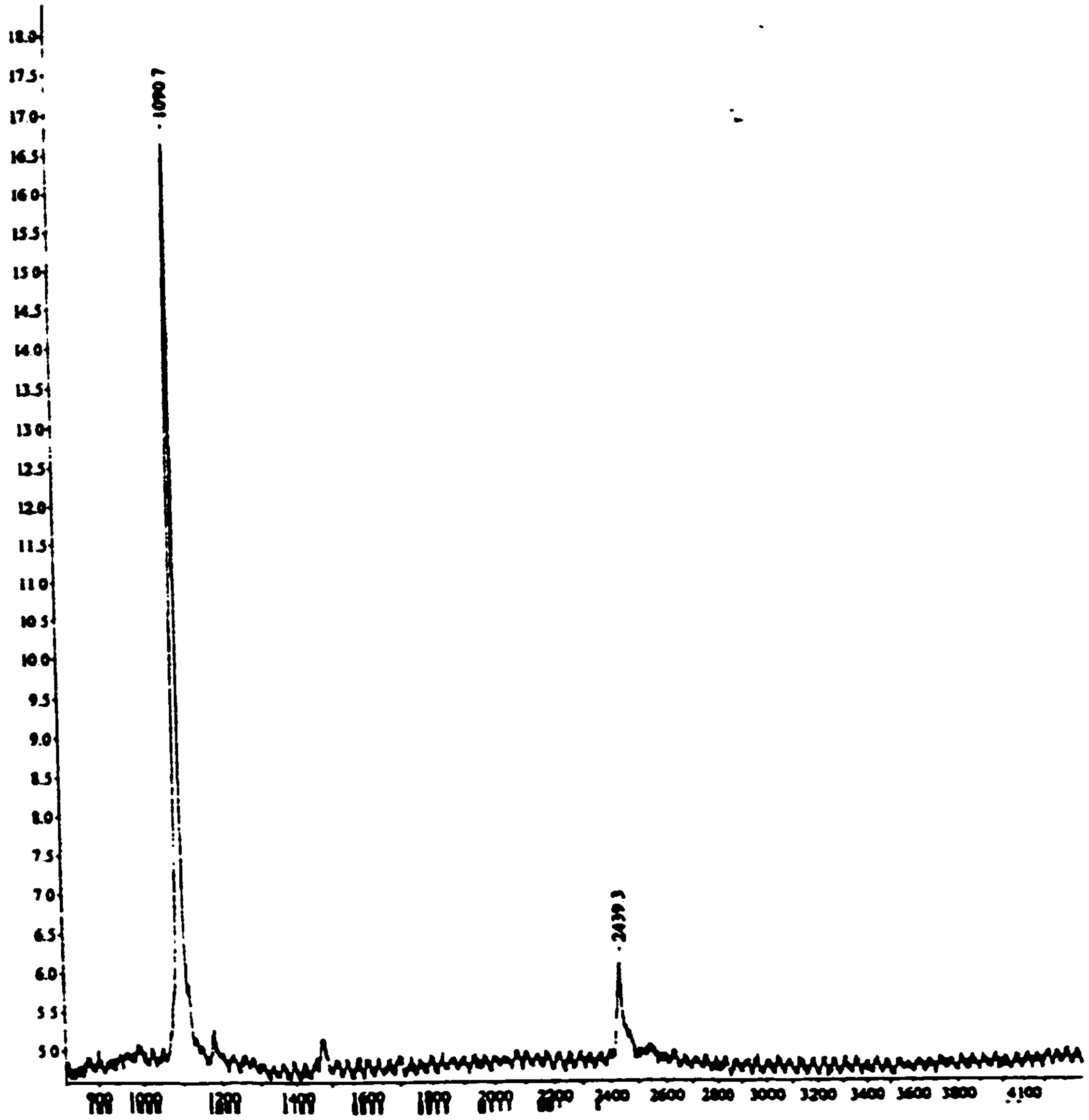

18.0
17.5
17.0
16.5
16.0
15.5
15.0
14.5
14.0
13.5
13.0
12.5
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
1090.7
2439.3
900
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4100

Figure 4-10

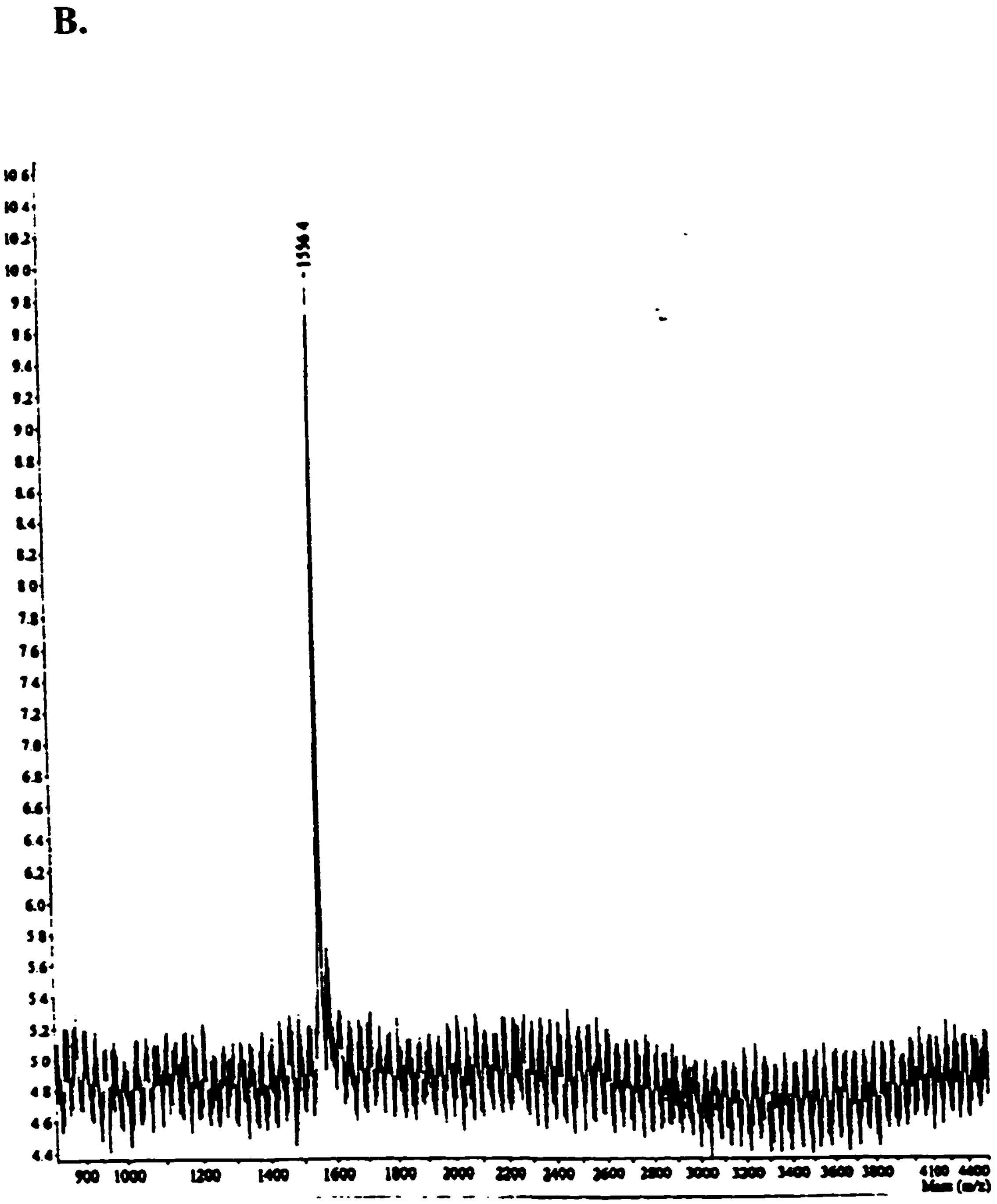

B.
1556.4
10.6
10.4
10.2
10.0
9.8
9.6
9.4
9.2
9.0
8.8
8.6
8.4
8.2
8.0
7.8
7.6
7.4
7.2
7.0
6.8
6.6
6.4
6.2
6.0
5.8
5.6
5.4
5.2
5.0
4.8
4.6
4.4
900
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4100
4400
Mass (m/z)

Figure 4-10

C.

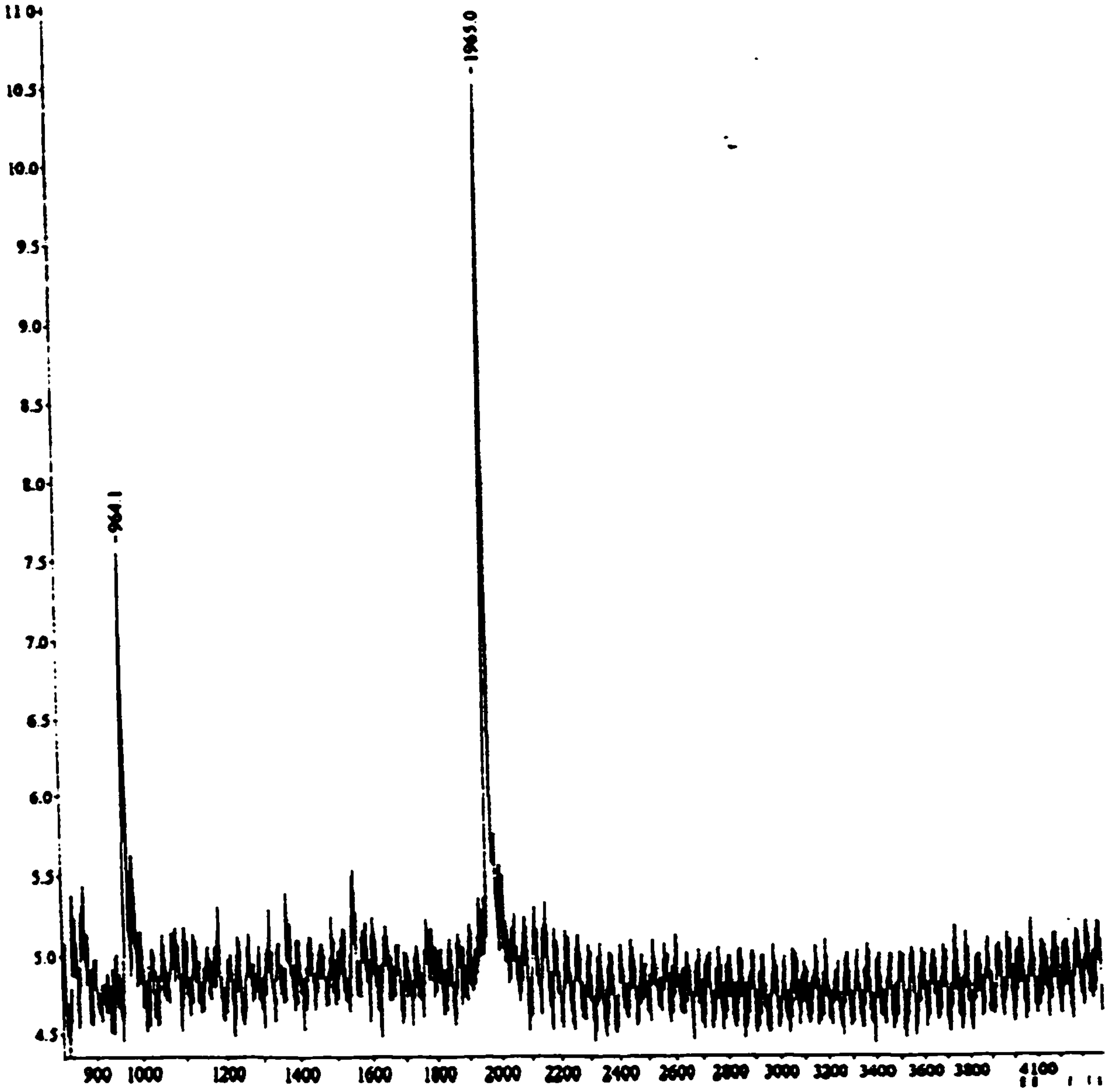
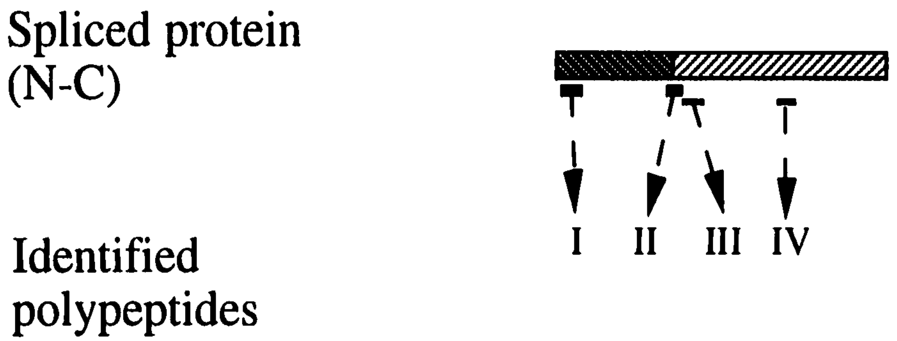

11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
964.1
1965.0
900
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4100

Figure 4-10

Figure 4-11 Identification of polypeptides of the spliced protein. Polypeptides II, III, and IV correspond to polypeptides #83, #112, and #107 in Fig. 4-9, respectively. Peptides I and II were identified by N-terminal sequencing and peptide sequencing, respectively. The determined sequences are shown (? marks undetermined residues). Peptides III and IV were identified by mass, with the measured value compared to predicted value.



I, sequence: ???KMDFLGLKNLTTLQRAV
II, sequence: FAEYCFNK
III, mass: measured 1965.0 / predicted 1967.2
IV, mass: measured 1556.4 / predicted 1555.8

Figure 4-11

additional *dnaE* gene (complete or in fragments), either. To confirm that this split *dnaE* gene is the only copy in the cyanobacterial cell, *Synechocystis* sp. PCC6803 total DNA was extracted as described in section 2.1.1.c. The total DNA was digested with various restriction endonucleases and analyzed on Southern blots. The 2.9-kbp probe (probe 1) and the-2.4 kbp probe (probe 2) contain the *dnaE-n* and *dnaE-c* genes, respectively. For each restriction enzyme that did not cut within the probe sequences, a single DNA band was detected (Fig. 4-12). A similar result was also observed in Southern blot analysis using the *E. coli dnaE* gene as probe (data not shown). Together, these results suggest that the split *dnaE* gene is the only copy of *dnaE* gene in this organism.

The study of the split *Ssp* DnaE intein in *E. coli* cells clearly shows that it is an active intein. However, the situation inside the cyanobacterial cell is not yet clear. Since the DnaE protein is essential for the cell, and no other *dnaE*-like gene was found in the genome, it is likely that the split DnaE intein is also active in the cyanobacterial cell. This possibility was examined *in vivo* by analyzing the protein products of the split *dnaE* gene in *Synechocystis* sp. PCC6803.

Synechocystis sp. PCC6803 cells were grown in liquid medium and harvested by centrifugation. The total cellular proteins were resolved on SDS-polyacrylamide gels and blotted onto nitrocellulose membrane. The membrane was then subjected to standard Western blot analysis using antibodies raised against the DnaE-n and DnaE-c proteins. However, no DnaE protein, either a spliced protein or precursors was detected (data not shown), probably due to the low level of the DnaE protein. To enrich the DnaE protein to a detectable level, Bio-Rex 70 column chromatography was used to isolate the DnaE protein. Bio-Rex 70 column is a cation exchange column and has been used in the purification of *E. coli* DnaE protein (Kim and McHenry, 1996a). Cells from approximately 4.5 liters of cyanobacterium culture were lysed by passage through a

Figure 4-12 Southern blot analysis of *Synechocystis* sp. PCC6803 total DNA. The positions of probe 1 and probe 2 are shown at the top. For lanes 1 to 8, the *Ssp* total DNAs were digested with restriction enzymes *Eco*RI, *Xba*I, *Hin*dIII, *Mfe*I, *Bam*HI, *Sma*I, *Bsp*EI, and *Afl*III, respectively. The expected sizes of the digestion products probed with probes 1 and 2 are listed below. Partial digestions were observed in lanes 2, 4 and 7.

Lane	**Enzyme**	**Expected size (bp):**	
		Probe 1	**Probe 2**
1	*Eco*RI	14,890	9,958
2	*Xba*I	7,924	N. D.
3	*Hin*dIII	14,118	5,031
4	*Mfe*I	4,445 + 2,946	N. D.
5	*Bam*HI	6,246	N. D.
6	*Sma*I	6,745*	N. D.
7	*Bsp*EI	N. D.	3, 785
8	*Afl*III	N. D.	N. D.

*: The expected size of the digestion products in lane 6 (probed with probe 1) is not consistent with the actual size, possibly due to a sequencing error in the genome sequence.

N. D: not determined.

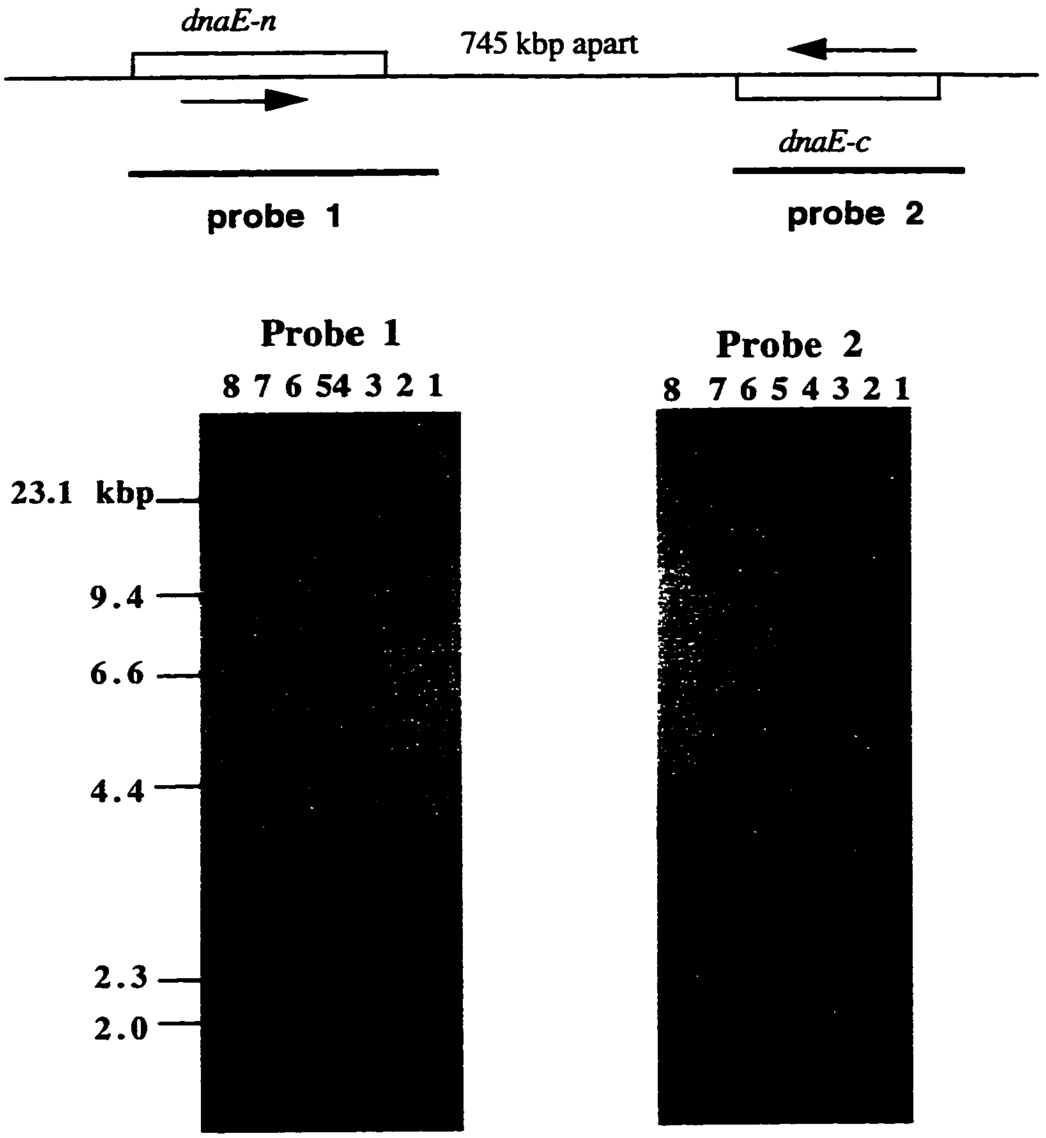

dnaE-n
745 kbp apart
dnaE-c
probe 1
probe 2
Probe 1
8 7 6 54 3 2 1
Probe 2
8 7 6 5 4 3 2 1
23.1 kbp
9.4
6.6
4.4
2.3
2.0

Figure 4-12

French pressure cell. The soluble proteins were then loaded on a 30-ml Bio-Rex 70 column. After washing away the unbound proteins, the bound proteins were eluted by 25 mM - 400 mM NaCl gradient. The fractions were concentrated by ~5 fold and subjected to Western blot analysis. In the fractions around 250 mM NaCl, Western blot analysis clearly showed two bands corresponding to the precursors, with the DnaE-n eluted in the earlier fractions and the DnaE-c protein in the later fractions (Fig. 4-13). A weak protein band with an estimated size of 137 kDa was also detected, with its apparent size matching the predicted size of the spliced protein. The fact that this protein can be specifically recognized by both anti-DnaE-n and anti-DnaE-c antibodies indicates that it is likely to be the spliced protein (ligated exteins). In the Western blot analysis using anti-DnaE-n antiserum, a smaller protein band was also recognized by the antiserum. It could be a break down product of the DnaE-n protein generated in the cyanobacterium cell, or during the purification process. These results indicate that the two *dnaE* genes are expressed in the cyanobacterial cells. An intact DnaE protein is likely produced at low level via protein *trans*-splicing.

Figure 4-13 Western blot analysis of DnaE protein of *Synechocystis* sp. PCC6803. The soluble fractions of *Ssp* total proteins were loaded onto a Bio-Rex 70 column. After washing away the unbound proteins, the bound proteins were eluted with a 25 mM to 400 mM NaCl gradient. The eluted fractions were subjected to Western blot using antisera raised against the DnaE-n and DnaE-c proteins. The positions of DnaE-n, DnaE-c and the 137-kDa protein are marked by arrows.

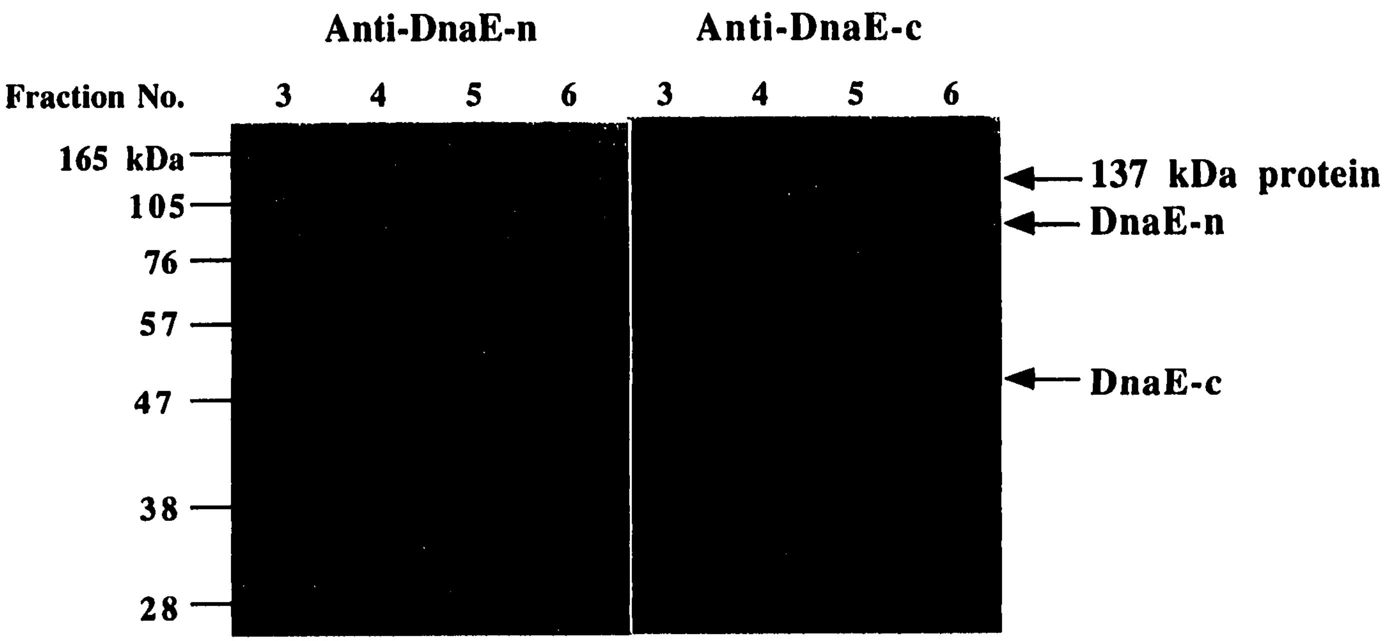

Anti-DnaE-n
Anti-DnaE-c
Fraction No.
3
4
5
6
3
4
5
6
165 kDa
105
76
57
47
38
28
137 kDa protein
DnaE-n
DnaE-c

Figure 4-13

4.3 Discussion:

4.3.1 The *Ssp dnaE* gene encodes a split intein capable of protein *trans*-splicing:

The *Ssp* DnaE intein was identified as a naturally occurring split mini-intein in *Synechocystis.* sp PCC6803, and was shown to be capable of protein *trans*-splicing. The two *dnaE*-like genes, *dnaE-n* and *dnaE-c*, are clearly two members of an intein-containing split *dnaE* gene, with the split being inside the intein coding sequence. Protein sequences deduced from the split *dnaE* gene, after excluding the intein sequences, reconstitute a complete DnaE protein that has neither a gap nor overlapping sequences at the split point. It also has the expected degrees of sequence identity to the continuous DnaE sequences of other bacterial organisms. The two intein sequences, Int-n and Int-c, not only have intein-like sequence features but also are proven to be two parts of a split intein by demonstrating a protein *trans*-splicing activity in *E. coli* cells. This *Ssp* DnaE intein, consisting of two separate polypeptides with a composite size of 159 aa, represents a split mini-intein that is apparently capable of forming a functional splicing domain through non-covalent interactions. Four conserved sequence motifs (A, B, F, G) have previously been localized in the splicing domain of inteins (Perler *et al.*, 1997; Duan *et al.*, 1997; Derbyshire *et al.*, 1997; Klabunde *et al.*, 1998; Pietrokovski, 1994; Shingledecker *et al.*, 1998), and all of them appear to exist in the *Ssp* DnaE intein, with motifs A and B located in Int-n and motifs F and G in Int-c (Fig. 4-4). The *Ssp* DnaE intein lacks a highly conserved His residue immediately before the C-terminal Asn. This His is replaced by an Ala. Four other inteins (*Ceu* ClpP, *Mja* PEP, *Mja* KlbA and *Mja* RpolA') also lack this penultimate His, where the His is replaced by Gly, Ser or Phe. This His has been shown to assist in Asn cyclization leading to cleavage of the peptide bond between the intein and C-extein (Xu and Perler, 1996), and efficient splicing of the *Ceu* ClpP intein in *E. coli* cells required a restoration of this His residue (Wang and Liu, 1997). The observation of *trans*-splicing activity with the *Ssp* DnaE intein shows that this His residue is not required for protein splicing of this intein. On the other hand, such a His residue need not necessarily be positioned next to the Asn in the primary amino acid sequence, and its function may be compensated by another His positioned close to the Asn in three-dimensional space but not in the primary sequence.

4.3.2 Protein *trans*-splicing of the split *Ssp* DnaE intein in cyanobacterial cells:

The *Ssp* DnaE intein very likely does protein *trans*-splicing in its native cyanobacterial cells, as it does in *E. coli* cells. A DnaE protein is essential for the cell, and there is no other *dnaE*-like gene (complete or partial) beside *dnaE-n* and *dnaE-c* in the *Synechocystis* sp PCC6803 genome. The detection of the two precursor proteins in the Bio-Rex 70 column-enriched proteins by Western blot analysis proved the expression of these two genes. This observation is consistent with the fact that the two genes maintain long open reading frames (2,694 bp for *dnaE-n* and 1,377 bp for *dnaE-c*), while their flanking sequences have numerous termination codons. Although the identity of the 137-kDa protein in the enriched proteins still needs further confirmation, it is likely to be a spliced protein. The production of a functional DnaE protein in the cyanobacterium cell most likely requires protein *trans*-splicing to remove the intein sequences and ligate the extein sequences. It is less likely, although possible, for the two precursor proteins (DnaE-n and DnaE-c) to reconstitute a functional protein without splicing, considering that the intein sequences interrupt both the β-binding domain and the τ–binding domain. Although the polymerase active site is contained entirely within the DnaE-n precursor protein, both the β–binding domain and the τ-binding domain are disrupted by the intein sequences and split between DnaE-n and DnaE-c precursor proteins. There is no indication that the half-intein sequences (Int-n and Int-c) can be cleaved off the precursor proteins without undergoing protein *trans*-splicing. Such a cleavage product was not observed with the DnaE-n and DnaE-c proteins in *E. coli*. Half-inteins engineered *in vitro* from other inteins also lack such a cleavage activity (see Chapter III; Southworth *et al.*, 1998; Shingledecker *et al.*, 1998). Functional β– and τ– binding domains are essential, because interactions of DnaE with the β subunit (DNA clamp) and the τ subunit are critical for the function of DNA polymerase III (Kim and McHenry, 1996).

A DnaE protein, either a spliced protein or precursors, could not be detected in the total protein of *Synechocystis* sp. PCC6803 by using the available anti-DnaE antisera without first enriching for the DnaE protein. This situation is most likely due to a combination of weak antisera and low levels of the DnaE protein. DnaE is known to exist at very low levels in other bacterial cells. The *E. coli* DnaE protein was estimated at 10-12 molecules per cell (Wu *et al.*, 1984), which is sufficient to replicate the *E. coli* genome approximately every 30 min. In comparison, *Synechocystis* sp. PCC6803 has a smaller genome that needs to be duplicated only every 10 hours (approximate cell doubling time). It is therefore not unreasonable for this organism to have extremely low levels of the DnaE protein for DNA replication.

4.3.3 Implications of finding of the split mini-intein on intein evolution and function:

The finding of a split mini-intein has implications for intein evolution. The *Ssp* DnaE intein could have evolved from a larger and continuous intein that later lost its sequence continuity. This alteration could occur through one or more genomic rearrangement events that separated the two halves of the *dnaE* gene (*dnaE-n* and *dnaE-c*) to different parts of the genome. A possible progenitor DnaE intein has not been found, and the 30% sequence identity between *Ssp* DnaE intein and the *Rma* DnaB intein (present in a DNA helicase) may be just coincidence, considering that the two inteins have non-homologous exteins and dissimilar insertion sites. Emergence of a split intein requires that it possesses protein *trans*-splicing activity, unless the exteins can function without ligation and without removal of the intein sequences. It has been reported recently that active *trans*-splicing elements can be reconstituted from separate intein fragments derived from inteins of continuous sequences (Southworth *et al.*, 1998;

Shingledecker *et al.*, 1998; Wu *et al.*, 1998a, also see Chapter III), suggesting that these inteins also have a potential of becoming split inteins. The *Ssp* DnaE intein (in fragments) has a total size of a mini-intein (splicing domain only) and lacks any of the endonuclease sequence motifs. The *Ssp* DnaE intein, like other inteins lacking an endonuclease domain, may once have had and lost the endonuclease domain (Telenti *et al.*, 1997), or, alternatively it may never have acquired an endonuclease domain. The split site in the *Ssp* DnaE intein coincides with predicted endonuclease insertion site, indicating that this site of the intein is tolerant of both insertion and cleavage. If the *Ssp* DnaE intein once had and lost its endonuclease domain, this loss could have occurred before or after the loss of sequence continuity. An intein presumably loses the ability of intein homing once the endonuclease domain is lost. As for the *Ssp* DnaE intein, having the two intein fragments encoded on different parts of the genome would prevent intein homing even if the endonuclease domain were still present.

Protein *trans*-splicing has been demonstrated with engineered intein fragments *in vivo* and *in vitro* (Southworth *et al.*, 1998; Shingledecker *et al.*, 1998; Mills *et al.*, 1998; Wu *et al.*, 1998a, also see Chapter III) and has produced insights into the structural requirements for protein splicing. The discovery of the *Ssp* DnaE intein, a natural split intein that does protein *trans*-splicing, provides a new perspective on this phenomenon. In terms of structural requirements for protein splicing, the size and sequence of this naturally evolved split mini-intein are in close agreement with those of the smallest functional mini-inteins that have been engineered so far in the laboratory (Derbyshire *et al.*, 1997; Wu *et al.*, 1998a, also see Chapter III). In terms of possible biological function, the *trans*-splicing reaction between the DnaE-n and DnaE-c precursor proteins may present a step where synthesis of a functional DnaE protein can be regulated. The absence of the penultimate C-terminal His residue (replaced by Ala) in the *Ssp* DnaE intein, although not preventing protein *trans*-splicing, may slow down the splicing reaction, as is the case for other inteins (Shao *et al.*, 1996; Xu and Perler, 1996; Wang and Liu, 1997). A slow and regulated splicing step may be a mechanism for assuring very low levels of production of the mature DnaE protein. The β and τ subunits of DNA polymerase III bind strongly with the DnaE protein and may therefore affect the *trans*-splicing reaction by bringing together the two precursor peptides of DnaE. It is interesting that an intein (the *Ssp* DnaX intein) also exists in the τsubunit of this organism (Liu and Hu, 1997a), although the *Ssp* DnaX intein has a continuous sequence and is not specifically related to the *Ssp* DnaE intein in sequence and insertion site. The presence of an intein in the τ subunit could be a regulation step for the trans-splicing reaction. Therefore, the synthesis of DnaE protein in the cyanobacteria can be controlled at different levels.

APPENDIX

The work described in this part of my thesis is a part of the research work carried out during my graduate program. It characterizes a histone-like protein encoded by a chloroplast gene. Because this part of the work is unrelated to the other parts of my study, and the results have been published, I attach the reprint of this paper as an appendix, which was recommended by my graduate supervisory committee.

A chloroplast gene (*hlpA*) in the cryptomonad alga *Guillardia theta* potentially encodes a protein resembling the bacterial histone-like protein HU. This gene was cloned and overexpressed in *E. coli* cells, and the resulting protein product, HlpA protein, was purified and characterized *in vitro*. The HlpA protein was shown to bind DNA in a sequence-independent manner, supporting the identification of this protein as a chloroplast HU-like protein. In addition to exhibiting a general DNA-binding activity, the chloroplast HlpA protein also strongly facilitated cyclization of a short DNA fragment in the presence of T4 DNA ligase, indicating its ability to mediate very tight DNA curvatures. The identification of this protein shows that chloroplasts, unlike mitochondria, have retained the ancestral HU protein from a bacterial endosymbiont to the present day.

Plant Molecular Biology 34: 339–343, 1997.
© 1997 *Kluwer Academic Publishers. Printed in Belgium.*

Short communication

DNA binding and bending by a chloroplast-encoded HU-like protein overexpressed in *Escherichia coli*

Hong Wu and Xiang-Qin Liu*
*Biochemistry Department, Dalhousie University, Halifax, Nova Scotia, B3H 4H7, Canada (*author for correspondence)*

Received 21 November 1996; accepted in revised form 18 February 1997

Key words: chloroplast, DNA bending, DNA binding, HU-like protein

Abstract

The *Guillardia theta* chloroplast *hlpA* gene encodes a protein resembling bacterial histone-like protein HU. This gene was cloned and overexpressed in *Escherichia coli* cells, and the resulting protein product, HlpA, was purified and characterized *in vitro*. In addition to exhibiting a general DNA-binding activity, the chloroplast HlpA protein also strongly facilitated cyclization of a short DNA fragment in the presence of T4 DNA ligase, indicating its ability to mediate very tight DNA curvatures.

Chloroplast genomes are often observed as highly condensed discrete structures termed nucleoids that associate with thylakoid and envelope membranes [7, 12, 16, 19]. In addition to DNA and RNA, a large number of proteins were found in chloroplast nucleoids [1, 7, 11, 15, 19]. Many of these proteins are probably required for the membrane-associating property and the transcriptional activity of the nucleoids, while others may resemble bacterial histone-like proteins and participate in the higher order structure or packaging of chloroplast DNA. In fact a spinach chloroplast protein cross-reacted with antibody against the bacterial histone-like protein HU [2], although a firm identification of this and other chloroplast nucleoid-associated proteins depends on future determination of their sequences and their ability to organize DNA. A mitochondrial DNA-binding protein (HM) of yeast and human has been shown to wrap (condense) and bend DNA *in vitro* [3, 6]. Surprisingly, the mitochondrial HM protein shows no sequence similarity to the bacterial histone-like protein HU, but instead it has a high degree of sequence similarity to the nuclear high mobility group (HMG) proteins [4, 14]. Because mitochondria and chloroplasts originated from bacteria through endosymbiosis, these findings led to the interesting suggestion that the ancestral HU protein of mitochondria has been replaced by the functionally similar HM protein evolved from a nuclear-encoded HMG1-like protein [10].

Previously we found in the cryptomonad alga *Guillardia theta* (formerly *Cryptomonas*) a chloroplast gene (*hlpA*) that potentially encodes a protein with 37% sequence identity to the *E. coli* HU protein over a 97-residue sequence [18]. In the present study, we overexpressed this chloroplast *hlpA* gene in *E. coli* cells and purified its protein product (HlpA). The HlpA protein was shown to bind DNA in a sequence-independent manner, supporting the identification of this protein as a chloroplast HU-like protein. The HlpA protein also strongly facilitated cyclization of a short DNA fragment in the presence of T4 DNA ligase, indicating its ability to bend DNA. The firm identification of this HU-like protein also shows that chloroplasts, unlike mitochondria, have retained the ancestral HU protein from a bacterial endosymbiont to the present day.

Overproduction and purification of the chloroplast HlpA protein

The complete chloroplast *hlpA* gene of *Guillardia theta* was cloned into the expression plasmid vector pET-16b (Novagen, Madison, WI), placing the *hlpA* gene directly behind a T7 promoter and inside *E. coli* strain

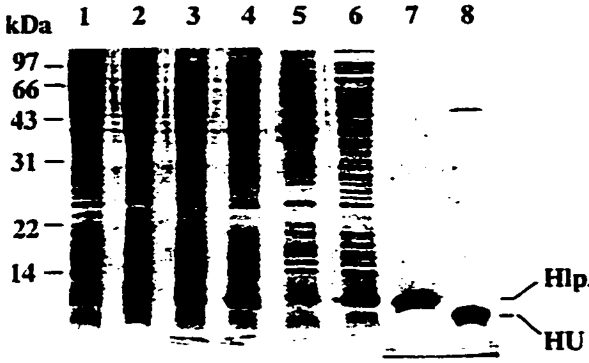


Figure 1. Overproduction and purification of the HlpA protein. All samples were electrophoresed on SDS-polyacrilamide gel and stained with Coomassie Blue. Lanes 1 and 2, total cellular protein from *E. coli* cells lacking the plasmid-borne *hlpA* gene before and after induction by IPTG, respectively. Lanes 3 and 4, total cellular protein from *E. coli* cells containing the plasmid-borne *hlpA* gene before and after induction by IPTG, respectively. Lanes 5 and 6, insoluble and soluble fractions of total protein seen in lane 4, respectively. Lane 7, purified HlpA protein. Lane 8, protein purified from *E. coli* cells lacking the plasmid-borne *hlpA* gene. The marking of a putative *E. coli* HU$_{\alpha\beta}$ protein (HU) is based on its apparent molecular mass of 9.5 kDa.

BL21(DE3) that harbors an IPTG-induced T7 RNA polymerase gene. Overproduction of HlpA protein was observed after a 3 h induction by IPTG, with the HlpA protein representing approximately 15% of the *E. coli* total protein (Figure 1, lane 4). The bulk of the overproduced HlpA protein was found in the soluble fraction of the cell lysate (Figure 1, lane 6). The HlpA protein showed strong binding on a DNA-cellulose column (Pharmacia) and was eluted off the column as a peak at about 500 mM NaCl. On a FPLC cation exchange column (Source 15S, Pharmacia), the HlpA protein was eluted also as a peak at about 500 mM NaCl. After these two steps of purification, the HlpA protein was apparently pure (Figure 1, lane 7), and no other protein band was detected on a heavily overloaded SDS-polyacrylamide gel by Coomassie Blue staining (data not shown). In a typical preparation, ca. 16 mg of HlpA protein was purified from 1 liter of the induced *E. coli* cell culture. As a mock control, *E. coli* cells lacking the recombinant chloroplast *hlpA* gene were subjected to the same protein induction and purification procedures as above. No protein was detected in elution fractions corresponding to that of the HlpA protein (data not shown). Instead, a protein of much lower quantity was eluted at about 430 mM NaCl on the FPLC cation exchange column. This protein resembled the *E. coli* HU$_{\alpha\beta}$ protein in electrophoretic mobility (apparent molecular mass 9.5 kDa) and migrated slightly faster than the HlpA protein (10.6 kDa) on a SDS-polyacrylamide gel (Figure 1, lane 8). The purified chloroplast HlpA protein was therefore judged to be free of significant contamination by the *E. coli* HU protein and other *E. coli* proteins.

DNA binding by the chloroplast HlpA protein

A gel shift assay was used to test whether the chloroplast HlpA protein has DNA binding activity. The DNA substrate was a 926-bp fragment cloned originally from the *Guillardia theta* chloroplast genome [18]. This DNA was first cleaved with restriction enzyme *Alu*I to produce a mixture of smaller DNA fragments ranging from 166 to 273 bp in size. As shown in Figure 2, the HlpA protein strongly retarded the gel mobility of all these DNA fragments, indicating that it binds DNA in a sequence-independent manner. Consistent with this notion, the HlpA protein also retarded gel mobility of DNA fragments of non-chloroplast origin and of different G+C content (data not shown). The DNA-HlpA complexes increased in size with increasing amounts of the HlpA protein, as indicated by the decreasing gel mobility of the complexes. At the largest HlpA amount tested (40 ng), the HlpA/DNA mass ratio was approximately 1, which translates into one HlpA molecule for approximately every 15 bp of DNA. Under this condition, the DNA-HlpA complexes were observed as multiple bands (Figure 2, lane 5) which are unlike the single broad band seen when lower amounts of HlpA protein were used (Figure 2, lanes 2–4). This may indicate that under this condition the DNA molecules were saturated with bound HlpA proteins, because the multiple bands represent different-sized DNA-HlpA complexes that appear to correlate well with the different-sized DNA fragments. As a mock control, samples prepared from *E. coli* cells lacking the recombinant chloroplast *hlpA* gene were also used in the DNA binding assay, and no detectable DNA-binding activity was detected in the FPLC elution fraction corresponding to elution position of the HlpA protein (Figure 2, lane 6), which again ruled out significant contamination of the purified HlpA protein by *E. coli* DNA-binding proteins.

DNA bending by the chloroplast HlpA protein

The chloroplast HlpA protein was also tested for its ability to mediate tight DNA curvatures (bending) by using the ring closure assay method. This method measures the acceleration of ligase-catalyzed cycliz-

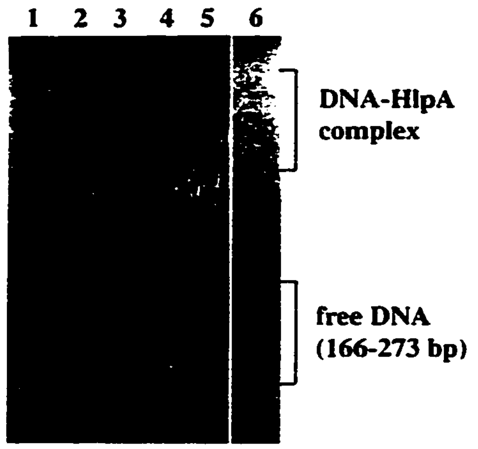


Figure 2. DNA binding by the HlpA protein. A mixture of [^{32}P]-labeled DNA fragments (38 ng) was incubated with purified HlpA protein, electrophoresed on a 7.5% polyacrylamide gel in 0.25× TBE buffer, and autoradiographed. Lane 1, no HlpA protein. Lanes 2, 3, 4 and 5 contained 16, 24, 32, and 40 ng of HlpA protein, respectively. Lane 6, mock control sample prepared from *E. coli* cells lacking the HlpA protein. The incubation was at 25 ° for 20 min in a 20 μl volume in 10 mM Tris-HCl pH 7.6, 15 mM KCl, 2 mM spermidine, 15% glycerol, 0.1 mM EDTA, and 0.1 mg/ml BSA.

ation of short DNA fragment and has been used previously in characterizing other DNA-bending proteins [8, 9]. Here a 117 bp DNA fragment with *Apa*LI cohesive ends was used as substrate in the ring closure assay. A DNA fragment of this short length is severely limited in its rate of covalent ring formation catalyzed by T4 DNA ligase, because the natural rigidity of the DNA double helix precludes close contact of the ends. In order to minimize the formation of multimers (linear or circular), a dilute solution (65 pM) of the 117 bp DNA was used in the ring closure assay. In the absence of the HlpA protein, the 117 bp linear monomer was not cyclized by T4 DNA ligase to a detectable level (Figure 3A, lane 2). Addition of the HlpA protein to the ligation reaction led to efficient cyclization of the linear monomer to form a circular monomer (Figure 3A, lanes 3–6).

The linear monomer and linear multimers in Figure 3A were readily identified by their sizes and sensitivity to exonuclease digestion. The circular monomer was identified by a combination of two tests. First, the circular monomer was resistant to degradation by λ exonuclease, while the linear monomer and linear multimers were readily degraded by the exonuclease (Figure 3A, lanes 7–9). The second test was to distinguish between circular monomer and circular multimer. As illustrated in Figure 3B, cleavage of a circular monomer by *Taq*I will produce only a 117 bp fragment, but cleavage of a circular multimer by *Taq*I will give rise to three fragments of 198 bp, 117 bp, and 36 bp in size. This is because a circular multimer (either dimer of larger) would be a mixture of isoforms produced through both head-to-tail and head-to-head ligations. When the ligation product was digested with *Taq*I, only a 117 bp fragment, but no 198 bp and 36 bp fragments, was produced from the circular molecule (Figure 3A, lane 10), which identified the circular molecule as a circular monomer rather than a circular multimer. A small amount of a 99 bp fragment was produced by *Taq*I cleavage of the unligated linear monomer remaining in the ligation product, while an expected 18 bp fragment ran off the gel. As a control, when the 117 bp DNA was ligated at a higher DNA concentration (3.25 nM), circular multimers were produced additional to circular monomer (Figure 3A, lanes 12 and 13). As predicted in Figures 3B, *Taq*I cleavage of these circular products produced a 198 bp fragment additional to the 117 bp fragment (Figures 3A, lane 14), while an expected 36 bp fragment migrated off the gel.

The chloroplast HlpA protein facilitated cyclization of the 117 bp DNA fragment most likely by inducing tight curvatures of the bound DNA molecule rather than by some other mechanisms, as has been discussed previously for similar ring closure assays [8, 9]. Although a precise mechanism for the HlpA-induced DNA bending is not known, it may be similar to a model proposed previously for the *E. coli* HU protein [17]. In this model, several wedge-shaped HU proteins bound adjacently along the DNA helix interact among themselves to close into a helical array with the DNA bound on the outside. Interestingly, when the HlpA/DNA molar ratio was approximately 2, the HlpA protein accelerated the formation of both circular monomer and linear multimers (Figure 3A, lane 3). When the HlpA:DNA molar ratio was increased to approximately 8 or higher, the circular monomer was formed almost exclusively (Figure 3A, lens 5 and 6). This observation may be explained by the above model, in which several HlpA proteins bound adjacently along a single DNA molecule are required to induce sufficient bending of the DNA for cyclization. When only one or two HlpA proteins are bound to each DNA molecule, the predominant interaction is between HlpA proteins bound to separate DNA molecules, which should facilitate the formation of linear multimers by bringing separate DNA molecules closer.

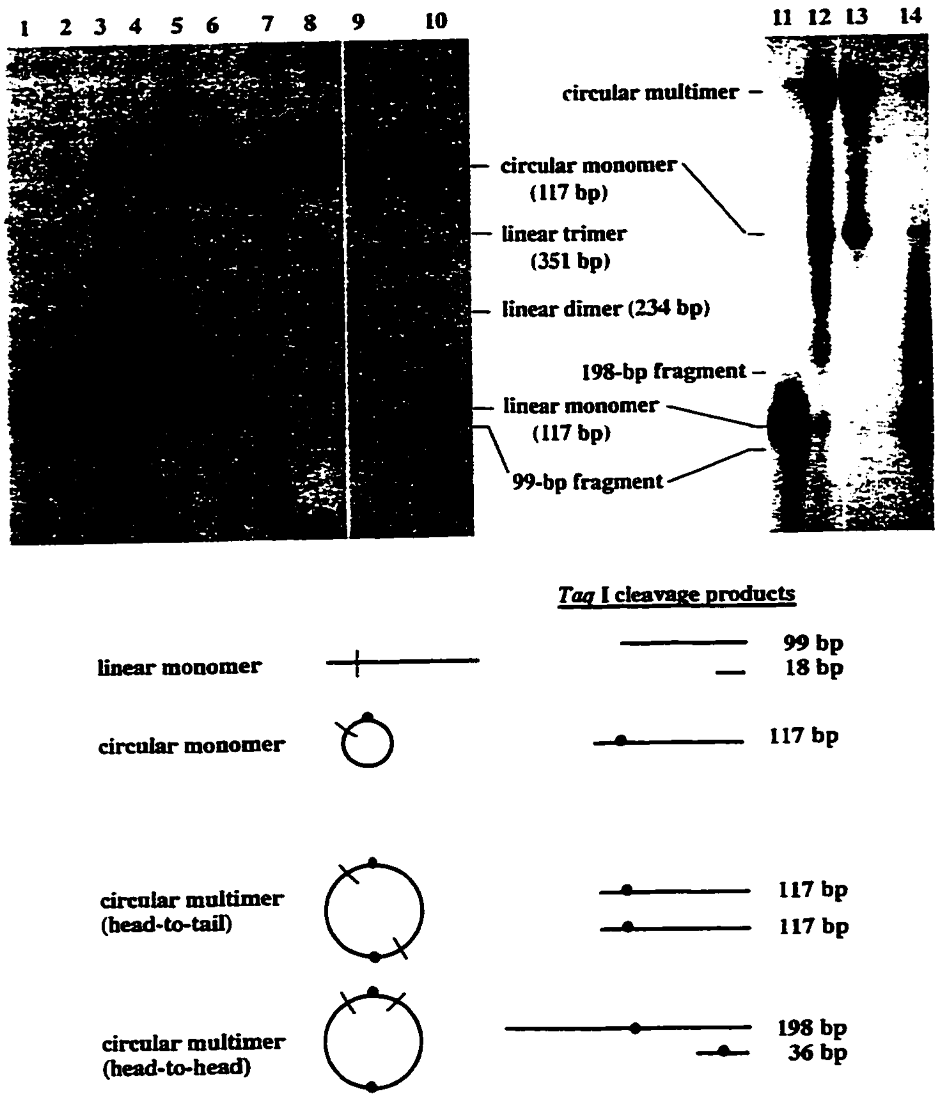


Figure 3. DNA bending by the HlpA protein. A (top). Ring closure assay. A [^{32}P]-labeled 117 bp DNA fragment with *Apa*LI cohesive ends was pre-incubated with purified HlpA protein at 25 °C for 30 min in a solution containing 5 ng/ml DNA, specified amount of HlpA, 50 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 1 mM ATP, 1 mM DTT, and 5% polyethylene glycol-8000. The DNA was then ligated by 40 units of T4 DNA ligase (25 °C, 2 h), deproteinized by phenol extraction, and electrophoresed on a 7.5% polyacrylamide gel in 0.25× TBE buffer. Lane 1, DNA before ligation. Lanes 2–6, ligation products after pre-incubation with HlpA at O (lane 2), 1.25 (lane 3), 2.5 (lane 4), 5.0 (lane 5), and 10 (lane 6) ng/ml. Lanes 7, 8 and 9 correspond to ligation products of lanes 3, 5, and 6, respectively, but after treatment with λ-exonuclease. Lane 10, ligation product of lane 6 treated with restriction enzyme *Taq*I. Lane 11, DNA before ligation. Lane 12, products of ligation at 250 ng/ml DNA and 200 ng/ml HlpA. Lanes 13 and 14, products of lane !2 treated with λ-exonuclease and *Taq*I, respectively. *B* (bottom). Schematic illustration of possible ligation products (left) and their corresponding *Taq*I cleavage products (right). DNA ligation sites are marked by a round dot. *Taq*I cleavage sites are marked by a short thin line.

Function and evolution of the chloroplast HlpA protein

The chloroplast HlpA protein is thus identified as an HU-like protein by its DNA-binding and DNA-bending activities in addition to its sequence similarity to the *E. coli* HU protein. To our knowledge, HlpA is the only chloroplast HU-like protein that has been identified and characterized by a combination of gene cloning, DNA binding, and DNA bending. Our observation that HlpA bends DNA *in vitro* suggests that it may function in chloroplast processes additional to DNA packaging.

By conferring flexibility (bending) to the bound DNA, the HlpA protein may serve as an accessory factor in stimulating other protein-DNA interactions. Like the *E. coli* HU and the mitochondrial HM proteins [5, 13], the chloroplast HlpA protein may play active roles in DNA replication, DNA recombination, and transcription. The availability of a purified HlpA protein and a cloned *hlpA* gene should facilitate the discovery of such functions.

It is interesting that the chloroplast HlpA protein resembles bacterial HU protein in sequence, in the light of mitochondrial HM protein whose sequence resembles the HMG1 family of nuclear non-histone proteins [4, 14]. Clearly the chloroplast HlpA protein is evolutionarily related to bacterial HU, the mitochondrial HM protein is evolutionarily related to nuclear HMG1, although these four proteins apparently are functional homologs [10, 13]. Thus chloroplasts and mitochondria have taken different evolutionary paths in this respect, with chloroplasts retaining the endosymbiont bacterial HU protein (HlpA) and mitochondria replacing HU with a homologue of the nuclear HMG1 protein (HM).

Acknowledgments

This work was supported by a research grant from the Natural Sciences and Engineering Research Council of Canada to X.-Q.L. who is a scholar in the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

References

1. Briat JF, Gigot C, Laulhere JP, Mache R: Visualization of a spinach plastid transcriptionally active DNA-protein complex in a highly condensed structure. Plant Physiol 69: 1205–1211 (1982).
2. Briat JF, Letoffe S, Mache R, Rouviere-Yaniv J: Similarity between the bacterial histone-like protien HU and a protein from spinach chloroplasts. FEBS Lett 172: 75–79 (1984).
3. Caron F, Jacq C, Rouviere-Yaniv J: Characterization of a histone-like protein extracted from yeast mitochondria. Proc Natl Acad Sci USA 76: 4265–4269 (1979).
4. Diffley JFX, Stillman B: A close relative of the nuclear, chromosomal high-mobility group protein HMG1 in yeast mitochondria. Proc Natl Acad Sci USA 88: 7864–7868 (1991).
5. Drlica K, Rouviere-Yaniv J: Histonelike proteins of bacteria. Microbiol Rev 51: 301–319 (1987).
6. Fisher RP, Lisowsky T, Parisi MA, Clayton DA: DNA wrapping and bending by a mitochondrial high mobility group-like transcriptional activator protein. J Biol Chem 267: 3358–3367 (1992).
7. Hansmann P, Falk H, Ronai K, Sitte P: Structure, composition, and distribution of plastid nucleoids in *Narcissus pseudomarcissus*. Planta 164: 459–472 (1985).
8. Hodges-Garcia Y, Hagerman PJ, Pettijohn DE: DNA ring closure mediated by protein HU. J Biol Chem 264: 14621–14623 (1989).
9. Laine B, Culard F, Maurizot J-C, Sautiere P: The chromosomal protein MC1 from the archaebacterium *Methanosarcina* sp. CHTI 55 induces DNA bending and supercoiling. Nucl Acids Res 19: 3041–3045 (1991).
10. Megraw TL, Chae C-B: Functional complementarity between the HMG1-like yeast mitochondrial histone HM and the bacterial histone-like protein HU. J Biol Chem 268: 12758–12763 (1993).
11. Nemoto Y, Kawano S, Kondoh K, Nagata T, Kuroiwa T: Studies on plastid-nuclei (nucleoids) in *Nicotiana tabacum* L. III. Isolation of chloroplast-nuclei from mesophyll protoplasts and identification of chloroplast DNA-binding proteins. Plant Cell Physiol 31: 767–776 (1990).
12. Nemoto Y, Kawano S, Nagata T, Kuroiwa T: Studies on plastid-nuclei (nucleoids) in *Nicotiana tabacum* L. IV. Association of chloroplast-DNA with proteins at several specific sites in isolated chloroplast-nuclei. Plant Cell Physiol 32: 131–141 (1991).
13. Oberto J, Drlica K, Rouviere-Yaniv J: Histone, HMG, HU, IHF: meme combat. Biochimie 76: 901–908 (1994).
14. Parisi MA, Clayton DA: Similarity of human mitochondrial transcription factor 1 to high mobility group proteins. Science 252: 965–969 (1991).
15. Reiss T, Link G: Characterization of transcriptionally active DNA-protein complexes from chloroplasts and etioplasts of mustard (*Sinapis alba* L.). Eur J Biochem 148: 207–212 (1985).
16. Sato N, Albrieux C, Yoyard J, Douce R, Kuroiwa T: Detection and characterization of a plastid envelope DNA-binding protein which may anchor plastid nucleoids. EMBO J 12: 555–561 (1993).
17. Tanaka I, Appelt K, Dijk J, White SW, Wilson KS: 3-Å resolution structure of a protein with histone-like properties in prokaryotes. Nature 310: 376–381 (1984).
18. Wang S, Liu X-Q: The plastid genome of *Cryptomonas* encodes an hsp70-like protein, a histone-like protein, and an acyl carrier protein. Proc Natl Acad Sci USA 88: 10783–10787 (1991).
19. Yurina NP, Belkina GG, Karapetyan NV, Odintsova MS: Nucleoids of pea chloroplasts: microscopic and chemical characterization. Occurrence of histone-like proteins. Biochem Mol Biol Int 36: 145–154 (1995).

REFERENCES

Barnes, M. H., and Brown, N. 1979. Antibody to *B. subtilis* DNA polymerase III: use in enzyme purification and examination of homology among replication-specific DNA polymerase. Nucleic Acids Research 6, 1203-1219.

Belfort, M. 1989. Bacteriophage introns: parasites within parasites? Trends in Genetics 5, 209-213.

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

Belfort, M., and Perlman, P. S. 1995. Mechanism of intron mobility. Journal of Biological Chemistry 270, 30237-30240.

Belfort, M., and Roberts, R. J. 1997. Homing endonucleases: keeping the house in order. Nucleic Acids Research 25, 3379-3388.

Bradford, M. M. 1976. A rapid and sensitive method fro the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Analytical Biochemistry 72, 248-254.

Bumcrot, D. A., Takada, R., and McMahon, A. P. 1995. Proteolytic processing yields two secreted form of sonic hedgehog. Molecular and Cellular Biology 15, 2294-2303.

Bult, C. J., White, W., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reich, C. J., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H.-P., Fraser, C. M., Smith, H. O., Woese, C. R., and Venter, J. C. 1996. Complete genome sequence of the Methanogenic archaeon, *Methanococcus jannaschii.* Science 273, 1058-1073.

Cadwell, R. C., and Joyce, G. F. 1992. Randomization of genes by PCR mutagenesis. PCR Methods and Applications 2, 28-33.

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

Chong, S., Shao, Y., Paulus, H., Benner, J., Perler, F. B., and Xu, M. Q. 1996. Protein splicing involving the *Saccharomyces cerevisiae* VMA intein: The steps in the splicing pathway, side reactions leading to protein cleavage and establishment of an *in vitro* splicing system. Journal of Biological Chemistry 271, 22159-22168.

Chong, S., and Xu, M. Q. 1997. Protein splicing of the *Saccharomyces cerevisiae* VMA intein without the endonuclease motifs. Journal of Biological Chemistry 272, 15587-15590.

Chong, S., Mersha, F. B., Comb, D. G., Scott, M. E., Landry, D., Vence, L. M., Perler, F. B., Benner, J., Kucera, R. B., Hirvonen, C. A., Pelletier, J. J., Paulus, H., and Xu, M. Q. 1997. Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. Gene 192:271-281.

Chong, S., Williams, K. S., Wotkowicz, C., and Xu, M. Q. 1998. Modulation of protein splicing of the *Saccharomyces cerevisiae* vacuolor membrane ATPase intein. Journal of Biological Chemistry 273, 10567-10577.

Clark, N. D. 1994. A proposed mechanism for the self-splicing of proteins. Proc. Natl. Acad. Sci. U. S. A. 91, 11084-11088.

Colleaux, L., d'Auriol, L., Betermier, M., Cottarel, G., Jacquier, A., Galibert, F., and Dujon, B. 1986. Universal code equivalent of a yeast mitochondrial intron reading frame is expressed into *E. coli* as a specific double strand endonuclease. Cell 44, 521-533.

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

Cooper, A. A. and Stevens, T. H. 1995. Protein splicing: Self-splicing of genetically mobile elements at the protein level. Trends in Biochemical Sciences 20, 351-356.

Dalgaard, J. Z., Moser, M. J., Hughey, R., and Mian, I. S. 1997a. Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. Journal of Computational Biology 4, 193-214.

Dalgaard, J. Z., Klar, A. J., Moser, M. J., Holley, W. R., Chatterjee, A., and Mian, I. S. 1997b. Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. Nucleic Acids Research 25, 4626-4638.

Davis, E. O., Sedgwick, S. G., and 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., Jenner, P. J., Brooks, P. C. Colsten, M. J., and Sedgwick, S. G. 1992.

Protein splicing in the maturation of *M. tuberculosis* RecA protein. Cell 71, 201-210.

Davis, E. O., Thangaraj, J. S., Brooks, P. C., and Colsten, M. J. 1994. Evidence of selection for protein intron in the RecAs of pathogenic *Mycobacteria*. EMBO Journal 13, 699-703.

Derbyshire, V., Wood, D. W., Wu, W., Dansereau, J. T., Dalgaard, J. Z., and Belfort, M. 1997. Genetic definition of a protein-splicing domain: Functional mini-inteins support structure predictions and a model for intein evolution. Proc. Natl. Acad. Sci. U. S. A. 94, 11466-11471.

Doolittle, R. F. 1993. The comings and goings of homing endonucleases and mobile introns. Proc. Natl. Acad. Sci. U. S. A. 90, 5379-5381.

Duan, X., Gimble, F. S., and Quiocho, F. A. 1997. Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. Cell 89, 555-564.

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

Fsihi, H., Vincent, V., and Cole, S. T. 1996. Homing events in the *gyrA* gene of some mycobacteria. Proc. Natl. Acad. Sci. U. S. A. 93, 3410-3415.

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

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

Gimble, F. S., and Thorner, J. 1993. Purification and characterization of VDE, a site-specific endonuclease from the yeast *Saccharomyces cerevisiae*. Journal of Biological Chemistry 268, 21844-21853.

Gimble, F. S., and Stephens, B. W. 1995. Substitutions in conserved dodecapeptide motifs that uncouple the DNA binding and DNA cleavage activities of PI-SceI endonuclease. Journal of Biological Chemistry 270, 5849-5856.

Gimble, F. S., and Wang, J. 1996. Substrate recognition and induced DNA distortion by the PI-SceI endonuclease, an enzyme generated by protein splicing. Journal of Molecular Biology 263, 163-180.

Gorbalenya, A. E. 1998. Non-canonical inteins. Nucleic Acids Research 26. 1741-1748

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

Hall. T. M. T., Porter, J. A., Young, K. E., Koonin, E. V., Beachy, P. A., and Leahy, D. J. 1997. Crystal structure of a hedgehog autoprocessing domains: Conservation of structure, sequence and cleavage mechanism between hedgehog and self-splicing proteins. Cell 91, 85-97.

Hammerschmidt, M., Brook, A., and McMahon, A. P. 1997. The world according to hedgehog. Trends in Genetics 13, 14-21.

He, Z., Crist, M., Yen, H., Duan, X., Quiocho, F. A., and Gimble, F. S. 1998. Amino acid residues in both the protein splicing and endonuclease domains of the PI-SceI intein mediate DNA binding. Journal of Biological Chemistry 273, 4607-4615.

Heath, P. J., Stephens, K. M., Monnat, R. J., Jr., and Stoddard, B. L. 1997. The structure of I-CreI, a group I intron-encoded homing endonuclease. Nature Structural Biology 4, 468-476.

Hirata, R., Ohsumi, Y., Nakano, A., Kawasaki, H., Suzuki, K., and Anraku, Y. 1990. Molecular structure of a gene, *VMA1*, encoding the catalytic subunit of H^+-translocating adenosine triphosphytase from vacuolar membranes of *Saccharomyces cerevisiae*. Journal of Biological Chemistry 265, 6726-6733.

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

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

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., and Tabata, S. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp strain PCC6803. DNA Research 3, 109-136.

Kawasaki, M., Nogami, S., Satow, Y., Ohya, Y., and Anraku, Y. 1997A. Identification of three core regions essential for protein splicing of the yeast Vma1 protozyme. Journal of Biological Chemistry 272, 15668-15674.

Keeling, P. J., and Roger, A. J. 1995. The selfish pursuit of sex. Nature 375, 283.

Kim, D. R., and McHenry, C. S. 1996a. *In vivo* assembly of overproduced DNA polymerase III: overproduction, purification, and characterization of the α, α-ε, and α-ε-θ subunits. Journal of Biological Science 271, 20681-20689.

Kim, D. R. and McHenry, C. S. 1996b. Identification of the β-binding domain of the α subunit of *Escherichia coli* polymerase III holoenzyme. Journal of Biological Chemistry 271, 20699-20704.

Klabunde, T., Sharma, S., Telenti, A., Jacobs, W. R., and Sacchettini, J. C. 1998. Crystal structure of gyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing. Nature Structural Biology 5, 31-36.

Koonin, E. V. 1995. A protein splice-junction motif in hedgehog family proteins. Trends in Biochemical Sciences 20, 41-42.

Kornberg, A. 1980. DNA replication. W. H. Freeman and Co., San Francisco.

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

Lambowitz, A. M., and Belfort, M. 1993. Introns as mobile genetic elements. Annual Review of Biochemistry 62, 587-622.

Lee, J. J., Effer, S. C., von Kessler, D. P., Porter, J. A., Sun, B. I., and Beachy, P. A. 1994. Autoproteolysis in hedgehog protein biogenesis. Science 266, 1528-1537.

Liu, X. Q., and Hu, Z. M. 1997a. Identification and characterization of a cyanobacterial DnaX intein. FEBS Letters 408, 311-314.

Liu, X. Q., and Hu, Z. M. 1997b. A DnaB intein in *Rhodothermus marinus*: indication of recent intein homing across remotely related organism. Proc. Natl. Acad. Sci. U. S. A. 94, 7851-7856.

Loizos, N., Tillier, E. R. M., and Belfort, M. 1994. Evolution of mobile group I introns: recognition of intron sequences by an intron-encoded endonuclease. Proc. Natl. Acad. Sci. U. S. A. 91, 11983-11987.

Michel, F., and Dujon, B. 1986. Genetic exchanges between bacteriophage T4 and Filamentous fungi? Cell 46, 323

Mills, K. V., Lew, B. M., Jiang, S., and Paulus, H. 1998. Protein splicing in *trans* by purified N- and C-terminal fragments of the *Mycobacterium tuberculosis* RecA intein. Proc. Natl. Acad. Sci. U. S. A. 95, 3543-3548.

Nakayama, N., Arai, N., Kaziro, Y., and Arai, K. 1984. Structural and functional studies of the dnaB protein using limited proteolysis: characterization of domains for DNA-dependent ATP hydrolysis and for protein association in the primosome. Journal of Biological Chemistry 259, 88-96.

Nogami, S., Satow, Y., Ohya, Y., and Anraku, Y. 1997. Probing novel elements for protein splicing in the yeast Vma1 protozyme: a study of replacement mutagenesis and intragenic suppression. Genetics 147, 73-85.

Perler, F. B. 1998. Protein splicing of inteins and hedgehog autoproteolysis: structure, function and evolution. Cell 92, 1-4.

Perler, F. B., Comb, D. G., Jack, W. E., Moran, L. S., Qiang, 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. U. S. A. 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., and Belfort, M. 1994. Protein splicing elements: inteins and exteins - a definition of terms and recommended nomenclature. Nucleic Acids Research 22, 1125-1127.

Perler, F. B., Olsen, G. J., and Adam, E. 1997. Compilation and analysis of intein sequences. Nucleic Acids Research 25, 1087-1093.

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

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.

Pietrokovski, S. 1996. A new intein in Cyanobacteria and its significance for the spread of inteins. Trends in Genetics 12, 287-288.

Pietrokovski, S. 1998. Modular organization of inteins and C-terminal autocatalytic domains. Protein Science 7, 64-71.

Porter, J. A., von Kessler, D. P., Ekker, S. C., Young, K. E., Lee, J. J., Moses, K., and Beachy, P. A. 1995. The product of hedgehog autoproteolytic cleavage active in local and long-range signaling. Nature 374, 363-366.

Porter, J. A., Ekker, S. C., Park, W.-J., von Kessler, D. P., Young, K. E., Chen, C.-H., Ma, Y., Woods, A. S., Cotter, R. J., Koonin, E. V., and Beachy, P. A. 1996a. Hedgehog patterning activity: role of a lipophilic modification mediated by the carboxyl-terminal autoprocessing domain. Cell 86, 21-34.

Porter, J. A., Young, K. E., and Beachy, P. A. 1996b. Cholesterol modification of hedgehog signaling proteins in animal development. Science 274, 255-259.

Reith, M. E., and Munholland, J. 1995. Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Molecular Biological Rep. 1995. 13, 333-335.

Shao, Y., Xu, M. Q., and Paulus, H. 1995. Protein splicing: characterization of the aminosuccinimide residue at the carboxyl terminus of the excised intervening sequence. Biochemistry 34, 10844-10850.

Shao, Y., Xu, M. Q., and Paulus, H. 1996. Protein splicing: Evidence for an N-O acyl rearrangement as the initial step in the splicing process. Biochemistry 35, 3810-3815.

Shingledecker, K., Jiang, S.-Q., and Paulus, H. 1998. Molecular dissection of the *Mycobacterium tuberculosis* RecA intein: design of a minimal intein and of a *trans*-splicing system involving two intein fragments. Gene 207, 187-195.

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

Southworth, M. W., Adam, E., Panne, D., Byer, R., Kautz, R. and Perler, F. B. 1998. Control of protein splicing by intein fragment reassembly. EMBO Journal 17, 918-926.

Telenti, A., Southworth, M. W., Alcaide, F., Daugelat, S., Jacobs, W. R., and Perler, F. B. 1997. The *Mycobacterium xenopi* GyrA protein splicing element: characterization of a minimal intein. Journal of Bacteriology 179, 6378-6382.

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

Wang, S. L., and Liu, X. Q. 1997. Identification of an unusual intein in chloroplast ClpP protease of *Chlamydomonas eugametos*. Journal of Biological Chemistry 272, 11869-11873.

Wechsler, J. W., and Gross, J. 1971. *Escherichia coli* mutants temperature sensitive for DNA synthesis. Molecular and General Genetics 113, 273-284.

Wende, W., Grindle, W., Christ, F., Pingoud, A., and Pingoud, V. 1996. Binding, bending and cleavage of DNA substrates by the homing endonuclease PI-SceI. Nucleic Acids Research 24, 4123-4132.

Wu, H., Xu, M. Q., and Liu, X. Q. 1998a. Protein *trans*-splicing and functional mini-inteins of a cyanobacterial DnaB intein. Biochimica et Biophysica Acta in press.

Wu, H., Hu, Z. M., and Liu, X. Q. 1998b. Protein *trans*-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803. Proc. Natl. Acad. Sci. U. S. A. 95, 9226-9231.

Wu, Y. H. Franden, M. A., Hawker, J. R., and McHenry, C. S. 1984. Monoclonal antibodies specific for the α subunit of the *Escherichia coli* DNA polymerase III holoenzyme. Journal of Biological Chemistry 259, 12117-12122.

Xu, M. Q., Southworth, M. W., Mersha, F. B., Horstra, 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.

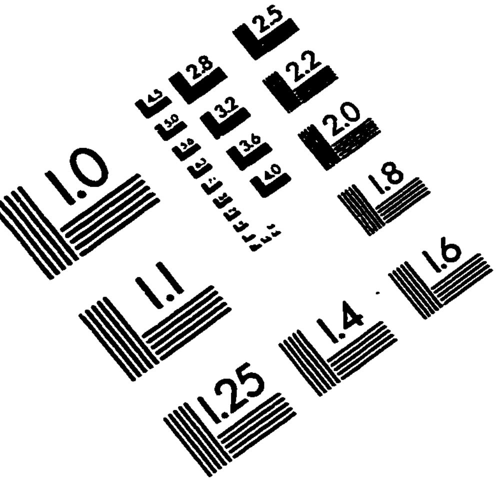
Xu, M. Q., Comb, D. G., Paulus, H., Noren, C. J., Shao, Y., and Perler, F. B. 1994. Protein splicing: an analysis of the branched intermediate and its resolution by succinimide formation. EMBO Journal 13, 5517-5522.

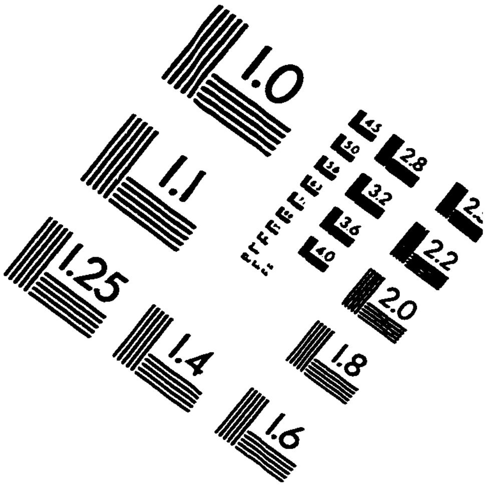
Xu, M. Q., and Perler, F. B. 1996. The mechanism of protein splicing and its modulation by mutation. EMBO Journal 15, 5146-5153.

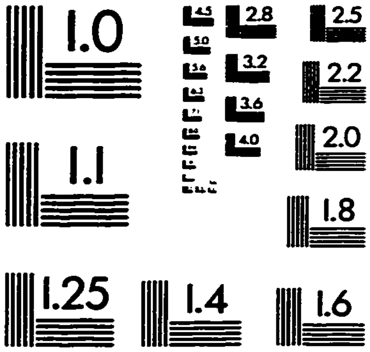
Xu, M. Q. 1997. The IMPACT of protein splicing research. The NEB Transcript 8, 1-5.

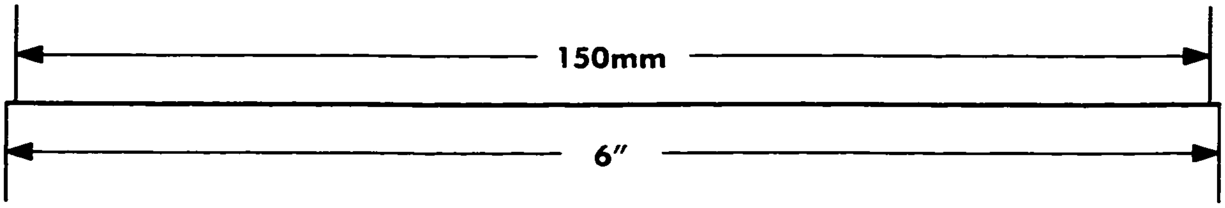
Zyskind, J. W., and Smith, D. W. 1977. *Escheria coli dnaB* mutant: direct involvement of the *dnaB252* gene product in the synthesis of an origin-ribonucleic acid species during initiation of a round of deoxyribonucleic acid replication. Journal of Bacteriology 125, 1476-1486.

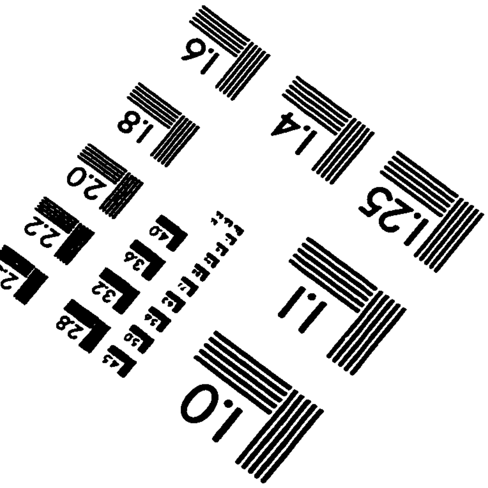
IMAGE EVALUATION TEST TARGET (QA-3)

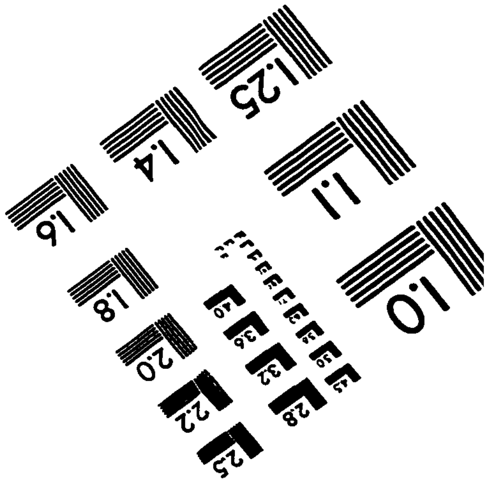












APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989