Structure of the chloroplast gene for the precursor of the Mr 32,000 photosystem II protein from mustard (Sinapis alba L.)

Gerhard Link and Ursula Langridge

Biologisches Institut II der Universität Freiburg, Schänzlestrasse 1, D-7800 Freiburg, FRG

Received 3 November 1983; Accepted 6 December 1983

ABSTRACT

The nucleotide sequence of the mustard chloroplast gene for the precursor of the Mr 32,000 photosystem II protein is presented. A comparison with the corresponding genes from spinach and Nicotiana debneyi (14) reveals less than 5% nucleotide divergence in the coding region. The derived protein of mustard differs from the corresponding proteins by three amino acid positions at the C-terminus. We have defined the presumed transcription start and termination sites of the mustard gene. Upstream from the start sites are sequences typical of a prokaryotic promoter and, also, a sequence that resembles the eukaryotic 'TATA' box. A search for intrastrand base pairing revealed stem-loop secondary structure at the transcription start and termination sites and in the region preceding the presumed promoter. This latter region is a 69-base-pair sequence element unique to the 5' flanking sequence of the mustard gene.

INTRODUCTION

Among the plant chloroplast proteins, a thylakoid membrane protein of Mr 32,000-35,000 has recently received much attention. This polypeptide is a major product of chloroplast protein synthesis but does not accumulate due to its rapid turnover (1,2). The protein was shown to be translated as a precursor of at least Mr 33,500-35,000 (3,4). Compelling evidence has been presented that the mature Mr 32,000 protein might be the primary target for triazine herbicides (5) and that its physiological function is the regulation of photosystem II electron transport (6,7). The single gene for the precursor of the Mr 32,000 photosystem II protein (pre-Mr 32,000 PSII protein) has been mapped on the circular chloroplast (cp)DNA molecule from a number of higher plants and algae (8-13). Nucleotide sequences have so far been published for the pre-Mr 32,000 protein genes from spinach and
Nicotiana debneyi (14).

It is well-established that the expression of the pre-M_{32,000} PSII gene is under light control (8, 15, 16). The mRNA was shown to be present in light-grown mustard seedlings in much higher levels than in dark-grown seedlings and evidence was obtained that the induction of this mRNA is mediated by the phytochrome (18) photoreceptor system (17). Since we are interested in studying transcriptional control mechanisms in mustard chloroplasts, it seemed appropriate to determine the nucleotide sequence of the coding region and of the 5' and 3' flanking regions of the PSII gene. With the sequence information presented in this paper we hope to provide a basis for investigations aimed at defining functionally the essential and regulatory elements for the transcription of this gene.

MATERIAL AND METHODS

Plasmids

Plasmids pSA364 and pSA452 together contain the entire pre-M_{32,000} PSII protein gene from mustard chloroplasts (11, 17). Plasmid pSA364a was constructed from pSA364 by subcloning the 1.1-kbp EcoRI-PstI border region of the insert. This intermediate plasmid was modified by deletion (19) of sequences adjacent to the EcoRI site, addition of a PstI linker (Boehringer Mannheim), and cloning of the resulting PstI insert in pBR322, giving rise to plasmid pSA364a-05. To facilitate sequence analysis, subclones of pSA364a-05 were constructed with pUC12 and pUC13 as vectors and JM83 as the host strain (20). These plasmids contain the 470-bp PstI(linker)-HincII segment (pSA05/A1), the 380-bp HincII-PstI segment (pSA05/A13), and the 330-bp PstI(linker)-SmaI segment (pSA05/D1)(Fig.1). Plasmid DNA was extracted from small liquid cultures by the rapid boiling method (21). For large scale isolation of plasmid DNA the alkaline extraction procedure (22) was used, followed by two cycles of CsCl/ethidium bromide density gradient centrifugations.

DNA sequencing

DNA fragments were labeled with $^{32}$P at their 5' ends, using T₄ polynucleotide kinase (BRL) and labeled ends were separated either following digestion with a second restriction enzyme or
by strand separation (23). Segments cloned into the pUC vectors were prepared for sequencing without prior gel purification. Linearized plasmids were cut with a second restriction enzyme and the resulting short oligonucleotide was removed by selective precipitation with isopropanol. The base-specific cleavage reactions were performed according to Maxam and Gilbert (23). Cleavage products were analyzed on 0.2 mm thin sequencing gels (24), covalently bound to one of the glass plates (25). Computerized nucleotide sequence analysis was performed, using the program described by Kröger and Kröger-Block (26).

**Exonuclease III/nuclease S1 digestion (19)**

Twenty μg of SmaI-digested plasmid pSA364a-O5 were incubated with sixty units E. coli exonuclease III (BRL) in 0.5 ml 70 mM Tris-HCl pH 8, 4 mM MgCl₂, 5 mM dithiothreitol at 30°C. Portions of 0.1 ml were withdrawn at 1-min intervals and immediately diluted into 0.5 ml 50 mM sodium acetate pH 4.5, 250 mM NaCl, 0.5 mM ZnSO₄. The DNA was incubated with 200 units nuclease S1 (Boehringer) at 30°C for 30 min, extracted with phenol and precipitated with ethanol. Samples were digested with PstI (BRL) and prepared for electrophoresis on a 7 M urea/6% acrylamide gel (27).

**Transcript mapping with nuclease S1 (28,29)**

For mapping of the 5'end of the transcript, the 5' labeled PstI insert of pSA364a-O5 (0.5 μg) was mixed with 30 μg of mustard chloroplast RNA (11) in 50 ul of 40 mM PIPES-NaOH pH 6.4, 400 mM NaCl, 1 mM EDTA, 80% formamide. The mixture was heated to 70°C for 10 min, then quickly transferred to 52°C and incubated for 5 hr, while the temperature was gradually lowered to 45°C. Upon chilling, the mixture was diluted tenfold with S1 digestion buffer (28) and was incubated with 100 units nuclease S1 at 37°C for 30 min. Following extraction with phenol and precipitation with ethanol, samples were resuspended either in 90% formamide for analysis on 7 M urea/6% polyacrylamide gels (27) or in 80% formamide, 10 mM NaOH, 1 mM EDTA for analysis on sequencing gels (24). Samples were heated at 90°C for 1 min, chilled, and loaded immediately onto the gel.

For mapping of the 3'end of the transcript, the 220-bp HindI-EcoRI fragment of the pSA452 insert was isolated and
labeled, using the Klenow fragment of *E. coli* DNA polymerase I (Boehringer) and alpha-^{32}P-dATP. Upon secondary digestion with TaqI (BRL) the 160-bp HinfI-TaqI fragment (0.2 µg) was isolated, hybridized with mustard chloroplast RNA, and treated with nuclease S1 as described above for 5' mapping.

RESULTS

Cloning and sequence analysis of the mustard PSII gene

The gene for the pre-M_{32,000} PSII protein is located in the large single-copy region of mustard cpDNA, near the border to one copy of the large inverted repeat (11,30). Plasmids pSA364 with cpDNA fragment Pst 5 as insert and pSA452 with fragment Pst 9 as insert together contain the entire coding sequence as well as extensive flanking regions of the PSII gene (11). Plasmid pSA364 contains at least two additional chloroplast genes of unknown function (G. Dietrich and G. Link, unpublished results). In an attempt to facilitate the sequence analysis of the PSII gene, a subclone pSA364a-05 was constructed from pSA364 as described in the 'METHODS' section. The 866-bp PstI insert of this subclone contains a portion of the PSII coding region and only several hundred base pairs of 5' flanking sequence. Fig. 1 shows a detailed map of this insert and of the 1.1-kb PstI-SmaI region of the pSA452 insert, which contains the major portion of the PSII gene and 3' flanking sequences. Fig. 1 also shows the sequencing strategy used for analysis of these DNA regions.

The nucleotide sequence of the region beginning at the SmaI site of the pSA364a-05 insert and ending near the TaqI site of the pSA452 fragment is shown in Fig. 2. The sequence contains an open reading frame of 1062 nucleotides, which codes for a 353-residue protein of approximately M_{39,000}. The derived protein is relatively rich in hydrophobic amino acids and contains no lysine.

Comparison of the nucleotide sequence of this open reading frame with the published sequences of the pre-M_{32,000} genes from spinach and *Nicotiana debneyi* (14) reveals that the three coding regions are identical in length and have more than 95% nucleotide positions in common (Fig. 2). Most of the 51 (spinach) or 52 (N. debneyi) nucleotide differences are silent with respect to
the amino acid sequence. However, there are three amino acids near the C-terminus which distinguish the derived mustard protein from the spinach and *N. debneyi* proteins: Isoleucine (residue 346), alanine (residue 348), and threonine (residue 351) in spinach and *N. debneyi* are changed to valine, serine, and isoleucine, respectively, in mustard (Fig. 2; codons at position +1136, +1142, and +1151).

The regions flanking the structural PSII sequences from mustard, spinach, and *N. debneyi* are conserved mainly within the 125 nucleotides immediately upstream from the presumed ATG translation start codon. Further upstream are numerous base exchanges and several deletions and insertions, the most notable being a 69-bp insertion at position -196 to -127 in the mustard sequence (Fig. 2). Inspection of this region reveals the presence of several sequences which could form stable stem-loop secondary structures (Fig. 3a). The presence of secondary structure in this region is supported by the results of nuclease digestion experiments (Fig. 3b). When plasmid pSA364a-05 was linearized at the single SmaI site (Fig. 1), digested with exonuclease III and nuclease S1, and then cut with PstI, the 330-bp PstI-SmaI

![Diagram](image-url)
fragment of the insert was successively shortened, depending on
the length of exonuclease III treatment. However, the 530-bp
SmaI-PstI fragment gave rise to a nuclease-protected fragment
approximately 440 bp in size. The border of this protected
region maps at position -180 (Fig. 2), i.e. within the 69-bp
insertion sequence.

Determination of the 5' and 3' ends of the PSII transcript

Hybridization of the 5'-labeled insert of pSA364a-O5 with
mustard cpRNA and treatment with nuclease S1 produced a single
S1-protected band of 340 nucleotides (Fig. 4a, lane 1). This
band was not generated in the control, in which the labeled PstI
insert was treated in the same way but in the absence of RNA
(lane 2). Since the PstI site at the left-hand border of the
insert had been constructed by adding a PstI linker (see METHODS),
the 5' label of the S1-protected fragment is located at the
(authentic) PstI site at the right-hand border of the insert
(Fig. 1). The 5' end of the in-vivo transcript thus maps at the
distance of approximately 340 nucleotides from this latter PstI
site. The direction of transcription is towards cpDNA fragment
Pst 9 (insert of pSA452), that is towards the large inverted
repeat in mustard cpDNA (11). There is no evidence of size
heterogeneity at the 5' end of the transcript. High-resolution
S1 mapping of the 5' end was carried out with the 85-bp Hinfl-
HincII fragment that spans the presumed transcription start site.
Fig. 4b shows the sizes of the S1-resistant products by calibra-
tion with a sequencing ladder derived from the same fragment.

Based on this analysis, the 5' end of the transcript is located
at or near position -86 (Fig. 2 and Fig. 4c). This region contains
sequences that are capable of forming the stem-loop structure

Fig. 2. Nucleotide sequence of the mustard PSII gene (M),
including flanking regions, in comparison with the PSII genes
(14) from spinach (S) and Nicotiana debneyi (N). The complete
mustard sequence is shown, while only those nucleotides that
differ in the spinach and N. debneyi sequences are given. ▲:
Single nucleotide deletions. ———: Deletions longer than two
nucleotides. The numbering relates to the mustard non-coding
strand sequence, in which +1 corresponds to the A of the first
ATG codon of the large open reading frame. Boxed codons: The
derived amino acids differ from those of the spinach and
N. debneyi sequences. Arrows: The apparent 5' and 3' ends of the
transcript as determined by S1 mapping.
Fig. 3. Secondary structure within the 69-bp insertion sequence in the 5' flanking region of the mustard PSII gene. (a): Proposed stem-loop structure as derived from the nucleotide sequence. (b): Exonuclease III/nuclease S1 digestion of the 5' portion and 5' flanking region of the mustard PSII gene. Digestion of pSA364a-05 with exonuclease III was for 1 min (lane 1), 2 min (lane 2), 3 min (lane 3), 4 min (lane 4), and 5 min (lane 5). Arrow: The 440-bp nuclease-resistant product from the 530-bp SmaI-PstI region. Markers: HinfI fragments of pBR322.

shown in Fig. 4c. The transcription start site is preceded by two sequence elements, GTTGACA and TGTTATAC, which resemble the consensus sequences for the '-35' (recognition) and '-10' (binding) region of prokaryotic promoters (34, 35).

The 3' end of the transcript was mapped with the 158-bp HinfI-TaqI fragment of the pSA452 insert. Sizes of the S1-resistant DNA bands were determined on a sequencing gel by comparison with the products of sequencing reactions generated from the same fragment (Fig. 5a). Two sets of S1-resistant bands are detected, which would correspond to the 3' ends of transcripts at positions +1117 and +1145 (Fig. 2). This region contains sequences which could produce the stem-loop structure shown in Fig. 5b. This structure is reminiscent of prokaryotic terminators (34).

DISCUSSION
Features of the coding region of the mustard PSII gene

The mustard chloroplast gene for the pre-M$_{r}$32,000 PSII
Fig. 4. Nucleotide sequence and proposed secondary structure of the DNA region at the 5'end of the mRNA. (a): S1 mapping of the 5'end of the mRNA on a 7 M urea/6% polyacrylamide gel. Arrows: Positions of the 866-bp insert of pSA364a-05 and the 340-ntd S1-resistant product. Lane 1: Products of S1 treatment following hybridization with chloroplast RNA. Lane 2: Products of S1 treatment following mock hybridization in the absence of RNA. Lane 3: HinfI fragments of pBR322 as size markers.

(b): High-resolution S1 mapping of the RNA 5'end on an 8% sequencing gel. The 85-bp HinfI-HincII fragment used was 5' labeled at the HinfI end. To measure the length of the S1-resistant products (S1), a sequence ladder from the coding strand of the same fragment was used for calibration. The relevant nucleotide sequence is written for both the coding strand (right) and the non-coding strand (left). The dot marks position -86 (Fig. 2).

(c): Nucleotide sequence of mustard cpDNA in the vicinity of the transcription start site. Arrows: Nucleotides at the 5'end of the mRNA as defined by S1 mapping. Underlined regions: Homology with the '-35' and '-10' regions of prokaryotic promoters (34,35). The proposed stem-loop structure containing the transcription start site is shown in the lower portion of the figure.

protein was previously mapped at the junction of cpDNA fragments Pst5 and Pst9 (11). We have now determined the nucleotide sequence of this gene and have compared it with the known (14) sequences of the PSII genes from spinach and N. debneyi. The coding region is of identical size in all three higher plant genes and has a high degree of sequence conservation. The derived amino acid sequence of the mustard PSII gene differs from the spinach and N. debneyi sequences in three positions near the
Fig. 5. Nucleotide sequence and secondary structure of the DNA region at the 3' end of the mRNA. (a): S1 mapping of the 3' end of the mRNA. The 158-bp HindIII-TaqI fragment of pSA452 was 3' labeled at the HindIII site. S1-resistant products (S1) were separated on an 8% sequencing gel alongside the products of sequencing reactions obtained from the same fragment. Two sets of samples were run for different times (left and right-hand panels). The nucleotides are given for both the coding strand (right) and the non-coding strand (left). (b): Proposed stem-loop structure at the 3' end of the mRNA. The derived structure of the non-coding DNA strand is presented. Arrows: Right-hand borders of the S1-resistant products at positions +1117 and +1145.

C-terminus, while these latter two sequences are entirely conserved over their full length (14). Cleavage at the C-terminus has been implicated in the precursor maturation of the Mr 32,000 PSII protein in Spirodela oligorhiza (J.B. Marder and M. Edelman, personal communication). This raises the possibility that the mature Mr 32,000 protein is completely conserved in mustard, spinach, and N. debneyi and that processing of the translation product involves cleavage at the less conserved C-terminus.

It is not known whether the N-terminus of the mature PSII protein or its precursor corresponds precisely to the presumed
ATG start codon of the mustard DNA sequence. The open reading frame codes for a predicted polypeptide of $M_r 39,000$, while the translation product of the mRNA has an apparent $M_r 35,000$ (11). A similar size discrepancy was noted for the PSII genes of spinach and *N. debneyi* (14) and was thought to reflect the unusual amino acid composition of the polypeptide, leading to anomalous electrophoretic mobility in SDS/polyacrylamide gels. Currently, one difficulty in defining the N-terminus is the lack of protein sequence data, due to the high turnover of the PSII protein (1,2). It is notable that the DNA sequence upstream from the ATG codon at position +1 (Fig. 2) does not reveal a possible ribosome binding site (31) complementary to the CCUCC motif of the 3' end of higher plant 16S rRNAs (32,33). A partial homology can, however, be deduced from the sequence preceding the second ATG codon at position +109. In spite of these uncertainties the authentic position of the translation start codon remains open for future investigation.

**Features of the flanking sequences**

Main aspects of the present work have been the precise map position and orientation of the PSII transcript, and the sequence elements that may be involved in the regulation of transcription. The 5' end of the mRNA is at or near position -86 of the nucleotide sequence (Fig. 2). This position is comparable to the 5' end of the mRNA in spinach (14). Furthermore, an *in-vitro*-transcription system from mustard chloroplasts synthesizes an RNA from the same 5' position (G. Link, unpublished results). The nucleotide sequence surrounding the presumed transcription start site of the mustard gene could form a short folded structure that may have a function in transcription, e.g. by directing the chloroplast RNA polymerase to the initiation site. An identical structure can be deduced from the spinach DNA sequence at the 5' end of the mRNA, while the corresponding structure in the *N. debneyi* sequence would be less stable due to a single C to A transversion in the stem region.

Sequence elements resembling the recognition and binding sites of prokaryotic promoters (34,35) occur in the mustard sequence immediately upstream of the mapped 5' end of the mRNA. Identical sequence elements are present in equivalent positions.
in the 5'flanking regions of the spinach and N. debneyi PSII genes (14). This 'prokaryotic' feature is shared with the putative promoter regions of other sequenced chloroplast genes (36). We note, however, that an additional conserved sequence element, TATATAAG, is present in the 5'flanking regions of all three PSII genes. This sequence is located 18 to 25 nucleotides upstream from the transcription start site (Fig. 2). It resembles the 'TATA' box of eukaryotic (RNA polymerase B or II) genes (37). However, since there is no sequence comparable to the 'CAAT' box of pol II genes further upstream in the 5'flanking region, this notion must await further experimental support. Likewise, the function of the 69-bp insertion sequence in the 5'flanking region of the mustard PSII gene is presently unknown. It could be that this highly structured region serves as an entry site for RNA polymerase or as binding site for an additional protein factor. However, there are no equivalent sequences in the flanking regions of the spinach and N. debneyi PSII genes. Further investigation of the 69-bp element may reveal if it is involved in modulating the transcription of the PSII gene.

There is little sequence homology among the 3'noncoding regions of the PSII genes from mustard, spinach, and N. debneyi. However, all three genes contain regions that can form a stem-loop structure similar to a prokaryotic terminator (34). When the mustard, spinach and N. debneyi sequences were aligned to maximize the homology (Fig. 2), these stem-loop structures map in an equivalent distance from the stop codon of the open reading frame. For the mustard gene, S1 mapping of the 3'end of the transcript was carried out (Fig. 5). Two sets of S1-resistant products were observed, of which the fragments ending near position +1117 could be due to reannealing of the labeled DNA strand in the stem region under the hybridization conditions used. Therefore, even though there may be two termination sites, it appears more likely that the S1-resistant fragments ending at or near position +1145 represent the authentic site. The resulting length of the 3'nontranslated region would be 82 nucleotides. The total length of the mRNA then becomes 1231 nucleotides, which agrees well with the size of 1.22 kb determined previously by Northern analysis (11).
In summary, the common features of the three PSII genes from mustard, spinach, and *N. debneyi* are their conserved coding region for the mature protein, start (and possibly stop) sites for transcription, and putative promoter elements. Other features, i.e. the derived C-terminal amino acid sequence and the 69-bp sequence element preceding the putative promoter region clearly distinguish the mustard gene from the spinach and *N. debneyi* genes. Studies on changes of these latter sequences, both naturally occurring and experimentally induced, and on the effect of such changes on the transcription of the PSII gene *in vivo* and *in vitro* can be expected to provide new insights in the mechanisms of (light-regulated) chloroplast gene expression.

ACKNOWLEDGEMENTS

We are grateful to Professor H. Mohr for encouragement and discussion; to Professor H. Kössel and Dr. P. Langridge for helpful suggestions and critical reading of the manuscript; to Dr. M. Kröger for help with the computer analysis; to Miss M. Krenz and Miss I. Roberson for expert technical assistance. Part of this work was performed during a visiting stay of one of the authors (G.L.) in Professor N.H. Chua's laboratory at the Rockefeller University, New York. This research was supported by the Deutsche Forschungsgemeinschaft to G.L. (Heisenberg-Programm/SFB206).

REFERENCES