Structural analysis of the short length ribosomal DNA variant from *Pisum sativum* L. cv. Alaska

Kenneth J.Piller, Scott R.Baerson, Neil O.Polans[1] and Lon S.Kaufman*
Laboratory for Cell, Molecular and Developmental Biology, Department of Biological Sciences, University of Illinois at Chicago, Chicago, IL 60680 and [1]Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115-2861, USA

Received March 27, 1990; Revised and Accepted April 17, 1990 EMBL accession no. X52575

ABSTRACT

The genomic clone, *RRN*pss1, representing the short ribosomal DNA length variant in *Pisum sativum* L. cv. Alaska, has been isolated and the 2859 bp intergenic spacer, along with the 25S rRNA 3′ border and 18S rRNA 5′ border, has been sequenced. The intergenic spacer contains nine tandem repeats, approximately 180 bp in length, which show greater than 80% sequence homology to each other. The RNA polymerase I transcription start site and a processing site, located 776 bp and 536 bp upstream of the 5′ end of 18S rRNA, respectively, have been determined using S1 analysis. The region surrounding the +1 site shows strong homology between the positions −6 to +10 to the rDNA sites of initiation in radish, maize, and wheat. The sequence CATGCAAA is located 19 bp upstream of the site of initiation, and appears once within each subrepeat and twice more between the end of the subrepeat array and the site of initiation. A previously characterized *Hpa*II site which shows developmental regulation of methylation is located 31 bp downstream of the site of initiation. Using RFLP linkage analysis, the short rDNA length variant of cv. Alaska is assigned to Chromosome 4 where it is genetically independent of the long rDNA length variant which is putatively assigned to Chromosome 7.

INTRODUCTION

Eukaryotic ribosomal RNA genes (rDNA) are arranged in tandem arrays consisting of 18S, 5.8S and 25S rRNA coding regions separated by internal transcribed spacers. These units in turn are separated by an intergenic spacer (IGS) extending from the 3′ end of the 25S rRNA coding region to the 5′ end of the 18S rRNA coding region. The IGS varies in size from a few kilobase pairs (kb) to more than 30 kb. This length heterogeneity is observed in both plants and animals, and can occur between and within species. Length heterogeneity within a species is generally due to variation in the number of copies of subrepeats (repetitive elements of approximately 100−400 bp) present within the IGS (for reviews, see 1,2,3).

The rDNA promoters of many animal systems, including human (4,5), mouse (6,7), rat (8), *Xenopus* (9,10), *Drosophila* (11) and *Acanthamoeba* (12,13) are well characterized. However, due to the lack of an adequate *in vitro* transcription system for plants, no plant rDNA promoters have been defined. Based on S1 data and sequence comparisons with published transcription start sites, putative sites of initiation have been identified in maize (14,15), wheat (16), and radish (17).

The IGS of most animal systems examined contains upstream control elements (1,3). These upstream control elements (UCEs), which can function as enhancers, occur either in the form of a single element, as in yeast (18), mouse (19), rat (20), and human (4), or in the form of subrepeats, as in *Xenopus* (21) and *Drosophila* (22,23). In all cases, except yeast, the UCEs contain sequences which are also found in the promoter. In all cases, except human, the UCEs are capable of initiating transcription. Developmental data from pea and genetic data from wheat indicate that the subrepeats in these systems may also function as UCEs (24).

Termination of rDNA transcription has been well studied in *Drosophila*, *Xenopus*, and mouse (reviewed in 25). In these organisms, a short sequence (9 bp to 18 bp), located downstream of the mature 25S rRNA 3′ end, directs termination of the primary transcript. In *Xenopus* (26) and mouse (27,28), copies of this sequence are located 215 bp and 171 bp upstream from the site of initiation, respectively. Deletion of these sequences results in a lower rate of transcript initiation. The manner in which these terminator sequences enhance initiation is unknown, although it has been shown that these terminator sequences can prevent promoter occlusion (29).

The rRNA genes of pea comprise two independent loci located on different chromosomes (30,31). Using RFLP analysis on a number of pea lines (excluding cv. Alaska), Polans et al. (31) have shown that the generally shorter, more homogenous length variants are located on the *Rrn1* locus of Chromosome 4 and the generally longer, sometimes heterogenous length variants are most likely located on the *Rrn2* locus of Chromosome 7, corresponding to the established locations of the pea nucleolar organizer regions (32,33,34). *Pisum sativum* L. cv. Alaska has two major length variants; a short length (8.6 kb) variant (S),

* To whom correspondence should be addressed at University of Illinois at Chicago, PO Box 4348, M/C 067, Chicago, IL 60680, USA

and a long length (9.0 kb) variant (L), presumably located at *Rrn1* and *Rrn2*, respectively. Several restriction fragment length polymorphisms (RFLPs) exist between the two variants, such as the presence of two *Eco*RI restriction sites within the array of subrepeats, and the absence of an *Sph*I site located at the 3′ end of the subrepeat array of the L variant (35).

Several lines of evidence point to differential regulation of the cv. Alaska length variants during light-mediated leaf development. Watson et al. (36) have examined cytosine methylation in the genomic rDNA during leaf development and have identified a *Hpa*II restriction site, near the promoter region of the S variant, whose methylation is developmentally regulated. Demethylation of the internal cytosine residue of the *Hpa*II recognition sequence CCGG is observed in the S variant as the lighting condition in the growth regime is changed from 7 days dark to 4 days dark followed by 3 days light; remethylation of this cytosine residue is observed when plants are grown in the light for 7 days (for growth regime conditions, see 37). The comparable cytosine residue in the L variant is unmethylated in all the above growth regimes.

Kaufman et al. (38) have determined that developmentally-regulated DNaseI hypersensitive sites are present near the promoter of the S variant. These sites are not present in the S variant when plants are grown for 7 days in the dark, but are observed in the S variant when plants are grown in the dark for 4 days followed by 3 days in the light. These sites are present in the L variant of plants grown in either condition.

Baerson and Kaufman (37) have determined that the rate of rDNA transcription is almost three times greater in nuclei isolated from plants grown in the dark for 4 days followed by 3 days in the light, than is the rate of rDNA transcription in nuclei isolated from plants grown in either 7 days darkness or 7 days light. These data, and the fact that nuclei obtained from light-grown plants contain more nucleoli per nucleus than do the nuclei from dark-grown plants (Kaufman, unpublished observation), suggest that the S variant is expressed only during certain stages of leaf development, while the L variant is expressed constitutively.

To aid in understanding the differential regulation of rRNA genes in pea, we have isolated and cloned the short rDNA length variant from *Pisum sativum* L. cv. Alaska. We report here the chromosomal location of this length variant, the complete nucleotide sequence of its IGS, and the possible significance of sequences and secondary structures located within the IGS.

MATERIALS AND METHODS

Plant material

Seed for inbred lines A73-91 and C879-344 were obtained through the courtesy of Dr. G. A. Marx (NYS Agr. Exp. Sta., Geneva, NY). *Pisum sativum* L. cv. Alaska seed was purchased from J. Mollema & Son, Inc. (Grand Rapids, MI).

Restriction endonuclease digestions and Southern analysis

Genomic DNA was digested overnight twice, using 6 U enzyme/μg of DNA on each occasion. Restricted DNA was fractionated in agarose gels (0.5% − 1.0%) and transferred to GeneScreen (NEN; Boston, MA) according to the manufacturer's guidelines. Blotting, hybridization, and washing conditions were performed as previously described (38).

For RFLP analysis, DNA was extracted and hybridized as previously described (31,39). Intergenic and coding regions were identified using the pea rDNA clone PHA1 (35) as previously described (31).

For copy number reconstruction, genomic DNA (5 μg) and pRRNpss1 DNA (66 ng, 120 ng, and 320 ng) was digested with *Hin*dIII, electrophoresed in a 0.5% agarose gel, blotted, and probed with nick-translated RRNpss1 insert. Relative hybridization was determined by using a GS300 Transmittance/Reflectance Scanning Densitometer (Hoefer Scientific Instruments; San Francisco, CA).

Genomic cloning and subcloning

A *Hin*dIII genomic library of *Pisum sativum* L. cv. Alaska in Lambda phage vector L47.1 (40) was obtained from the laboratory of W.F. Thompson (North Carolina State University, NC). Plaques were screened as described by Maniatis et al. (41) using the pea minor rDNA length variant, pHA1 (35), as a probe. Putative short rDNA length variant clones were identified via restriction site analysis and one, *RRN*pss1, was selected for further study (figure 1). The subclone pEB3890 was constructed by ligating the *Eco*RI/*Hin*dIII 2550bp fragment (containing the 3′ end of the 25S rDNA and upstream portion of the IGR) and the *Hin*dIII/*Bam*HI 1340bp fragment (containing the downstream portion of the IGR and 5′ end of the 18S rDNA) into the plasmid vector Bluescribe (Stratagene; La Jolla, CA) at the *Eco*RI and *Bam*HI sites of the polylinker.

DNA Sequencing

Sequencing was performed in both directions using the Sequenase kit (United States Biochemical; Cleveland, OH). Templates for sequencing were made by subcloning fragments of pEB3890 into the phage vectors M13mp18 and M13mp19 (42). Sequence specific primers (17 bp-long) were synthesized on a Biosearch 8700 DNA synthesizer (Biosearch, Inc.; San Rafael, CA). In cases where compressions occurred, the guanine analog 7-deaza dGTP was used to resolve the sequence (43). DNA samples were electrophoresed in 45cm − 85cm long polyacrylamide gels (4% − 9%) containing 8M urea. DNA sequence comparisons and primary and secondary structure analysis were carried out using previously detailed algorithims (44).

S1 nuclease mapping

Poly A$^-$ RNA was obtained by running total RNA isolated from the leaves of seven-day-old white-light-grown peas over a poly (U) Sephadex (BRL) column and collecting the effluent as previously described (45). The effluent was passed over the column a second time, and was again collected. For each S1 experiment, 300 μg of poly A$^-$ RNA was coprecipitated with 10^6 cpm of the appropriate 5′ end-labeled (HX194, PR249, XX663) or 3′ filled-in (HpBg311) probe. Labeling, annealing and S1 digestion were performed as in Maniatis et al. (41) except that buffer containing 120 units of S1 nuclease was added to each reaction. Conditions of S1 digestion differed for each probe and were as follows: 49°C for 15 minutes (HX1940), 42°C for 10 minutes (PR249), 49°C for 5 minutes (XX663), or 60°C for 10 minutes (HBg311). All reactions were phenol/chloroform extracted, ethanol precipitated, and dissolved in 15 μl of loading buffer (36% formamide, 8mM EDTA, 0.02% Bromophenol Blue, 0.02% Xylene Cyanol FF). A standard volume of 3 μl was loaded onto sequencing gels, along with sequencing reactions of M13mp18 which functioned as length markers.

The locations of DNA fragments used as probes for S1 analysis are shown in figure 1; HpBg311 is a 311 bp *Hpa*II/*Bgl*II

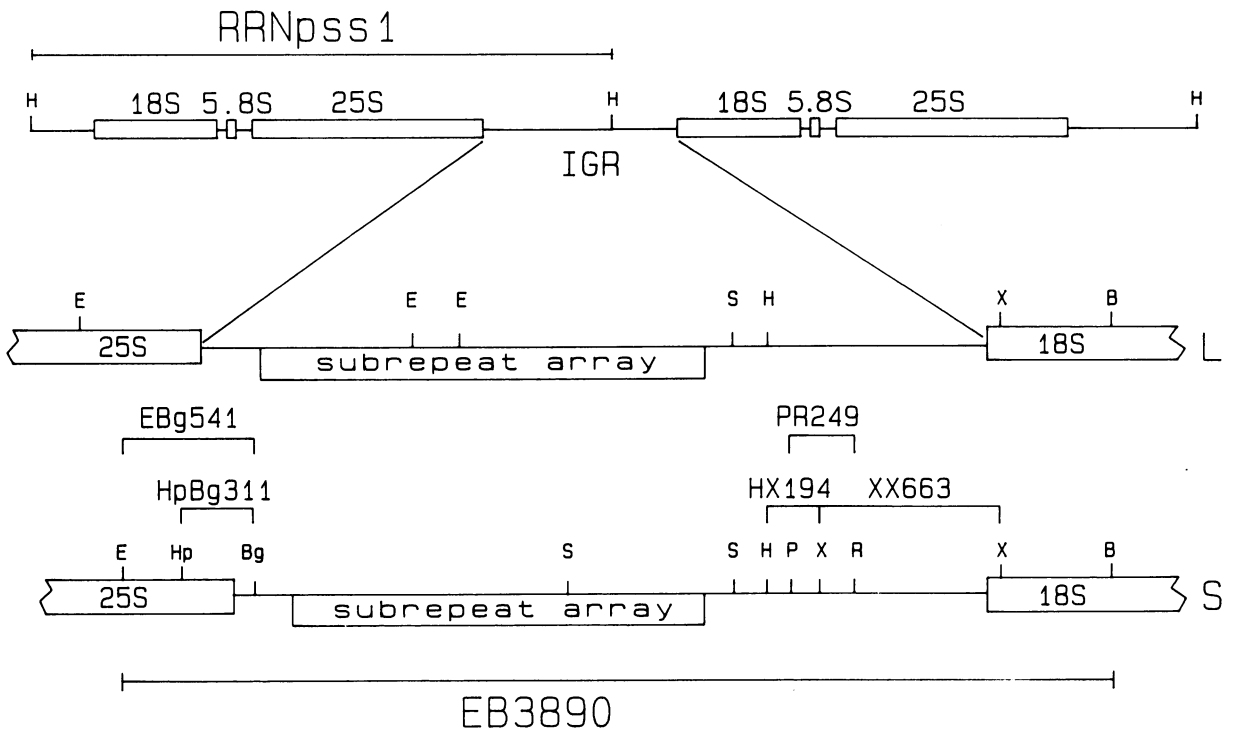

Figure 1. The major rDNA length variants in pea. The long (L) and short (S) variants differ in length (400 bp) and restriction site composition within the intergenic spacer (IGS). *RRN*pss1 is a genomic clone representative of the S variant family. EB3890 is a subclone of *RRN*pss1 containing the 25S rRNA 3′ border, IGS, and 18S rRNA 5′ border. Probes which generated S1 nuclease protected fragments are shown (HX194, PR249, XX663, HpBg311). The probe EBg541 was used for indirect end-labeling (38). The restriction sites *Hin*dIII (H), *Eco*RI (E), *Sph*I (S), *Bam*HI (B), *Xmn*I (X), *Hpa*II (Hp), *Bgl*II (Bg), *Pst*I (P) and *Rsa*I (R) are indicated.

fragment; HX194 is a 194 bp *Hin*dIII/*Xmn*I fragment; PR249 is a 249 bp *Pst*I/*Rsa*I fragment; XX663 is a 663 bp *Xmn*I/*Xmn*I fragment.

RFLP analysis of chromosomal location

F_2 progeny from crosses Alaska×A73-91 and C879-344 × Alaska were individually scored for isozymes, morphological traits, and rDNA variation as previously described (31,39). Linkage relationships were estimated using the 3.10 version of the LINKAGE-1 computer program (46).

RESULTS AND DISCUSSION

Chromosomal locations of the two major rDNA length variants in *Pisum sativum* L. cv. Alaska

The rRNA genes of pea are confined to two genetic loci, *Rrn1* and *Rrn2*, located on Chromosomes 4 and 7, respectively (31). In that study, which did not include cv. Alaska, digestion with *Eco*RI divided the pea rDNA repeats into a 3.7 kb fragment contained completely within the rRNA transcription unit, and a series of fragments (varying in length by multiples of 180 bp) containing part of the IGR. These latter bands were designated by Polans et al. (31) as follows: a (approximately 4.6 kb in length), b (approximately 4.8 kb in length), c (approximately 5.2 kb in length), and d (a series of six bands ranging from approximately 5.5 kb−10.0 kb in length). Pea inbred lines C879-344 and A73-91 display the bd and ac haplotypes, respectively. Genetic analysis of the F_2 progeny derived from their cross indicated that *Rrn1* corresponded to the shorter a and b bands, while *Rrn2* corresponded to the longer c and d bands. Using these and other crosses, *Rrn1* was mapped to Chromosome 4 because of the tight linkage it displayed with Chromosome 4 markers *Fa* (fasciated stem phenotype) and *Was* (a waxy mutant phenotype). *Rrn2* showed tight linkage with a group of unmapped isozyme loci and was putatively assigned to Chromosome 7 based upon both genetical and cytological arguments.

In this study, assignment of the two major rDNA length variants of cv. Alaska was accomplished by crossing Alaska with the two lines used in the previous study and then examining the segregation patterns that these crosses produced in the F_2 generation among the mapped and unassigned length variants. Both crosses were performed because the Alaska band pattern originally appeared to be bbcc, permitting segregation analysis for *Rrn1* (a,b) using only the Alaska (bbcc)×A73-91 (aacc) cross and for *Rrn2* (c,d) using only the the C879-344 (bbdd)×Alaska (bbcc) cross. However, due to the presence of two *Eco*RI restriction sites within the L variant subrepeat array of Alaska (see figures 1 and 3) and what appears to be differential methylation of these cleavage sites according to genetic background (Polans, unpublished observation), band c was often replaced by (or appeared with) its *Eco*RI digested products−most notably a 1.4 kb fragment denoted f. Thus, the second cross is more appropriately described as C879-344 (bbdd)×Alaska (bbff). In either case, the F2 banding patterns were unambiguously scored and the data summarized in Table 1. For *Rrn1*, *Rrn2* and *Fa* (the only relevant genetic marker segregating in these crosses), a comparison of observed and expected F_2 progeny in each category showed excellent agreement with the 1:2:1 Mendelian ratio for segregation at a single locus, as indicated by χ^2

Table 1. Individual and joint segregation analysis involving *Rrn1*, *Rrn2* and an associated marker locus.

Loci	F_2 progeny in each genotypic class[a], no. 11	12	22	11,11	11,12	11,22	22,11	22,12	22,22	n	χ^2	P
					ALASKA×A73-91							
Rrn1[b]	8	10	6							24	1.00	0.61 (N.S.)
Fa[c]	5		19							24	0.22	0.64 (N.S.)
Fa, *Rrn1*[d]				4	1	0	4	9	6	24	6.42	<0.05
					C879-344×ALASKA							
Rrn2[b]	17	24	15							56	1.29	0.53 (N.S.)

[a] Genotypic designations: 1 = A73−91 or C879−344 haplotype, 2 = Alaska haplotype
[b] 1:2:1 expected ratio (2 d.f.)
[c] 1:3 expected ratio (1 d.f.); heterozygote cannot be differentiated due to dominance
[d] 1:2:1:3:6:3 expected ratio (2 d.f.); recombinant fraction = 0.22 ± 0.09

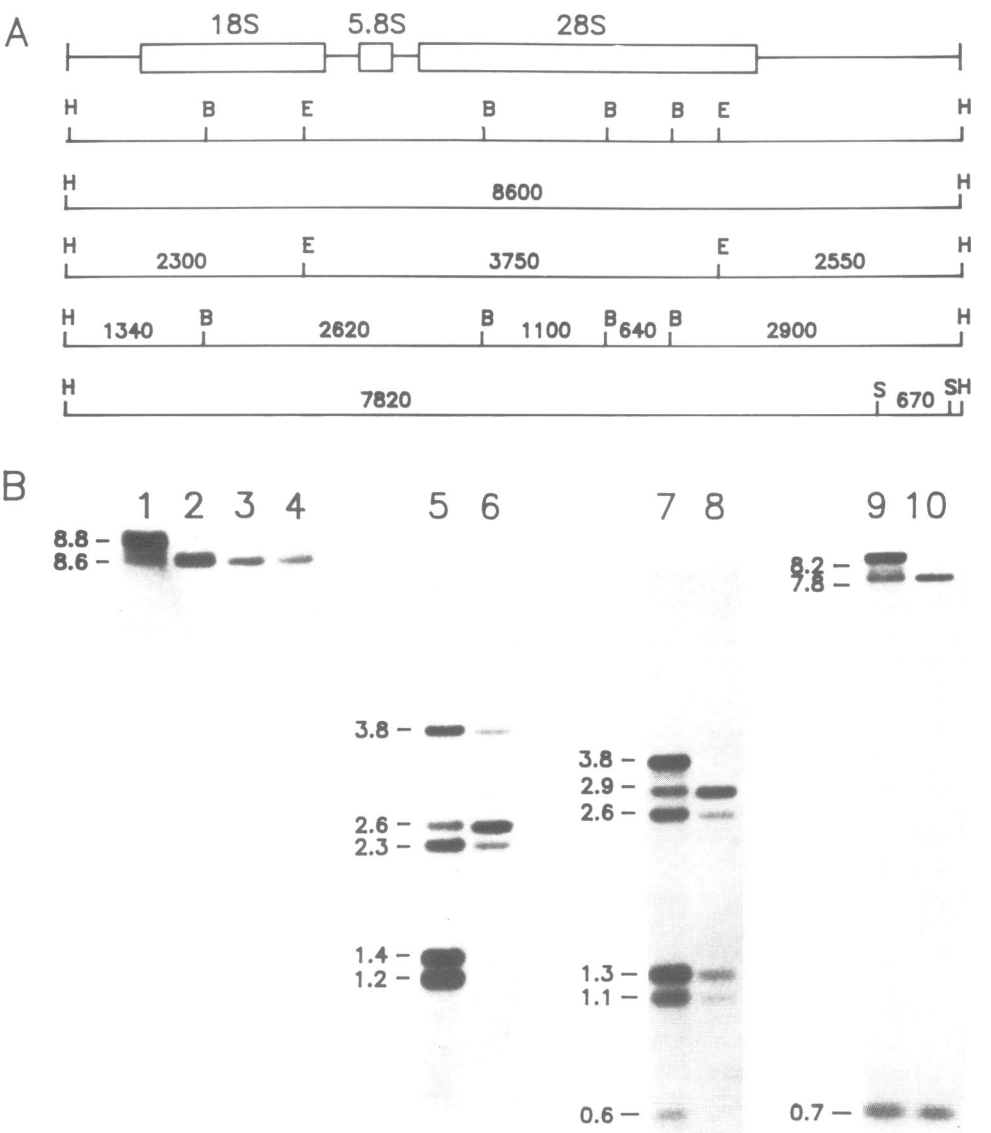

Figure 2. Restriction and Southern analysis of pRRNpss1. (A) A partial restriction map of *RRN*pss1 contains one *Hin*dIII site (H), two *Eco*RI sites (E), four *Bam*HI sites (B), and two *Sph*I sites (S). The sizes of generated fragments are given in base pairs. (B) Genomic DNA (lanes 1,5,7,9) and pRRNpss1 DNA (lanes 2,3,4,6,8,10) were digested with the following restriction enzymes: Lanes 1 thru 4, *Hin*dIII; lanes 5 and 6, *Hin*dIII plus *Eco*RI; lanes 7 and 8, *Hin*dIII plus *Bam*HI; lanes 9 and 10, *Hin*dIII plus *Sph*I. Nick-translated RRNpss1 insert was used as a probe. Copy number reconstruction (lanes 1 through 4) was performed using 5 μg of genomic DNA (lane 1) and 320 ng, 120 ng and 66 ng of pRRNpss1 DNA (lanes 2,3, and 4, respectively). Sizes of fragments are given in kilobase pairs.

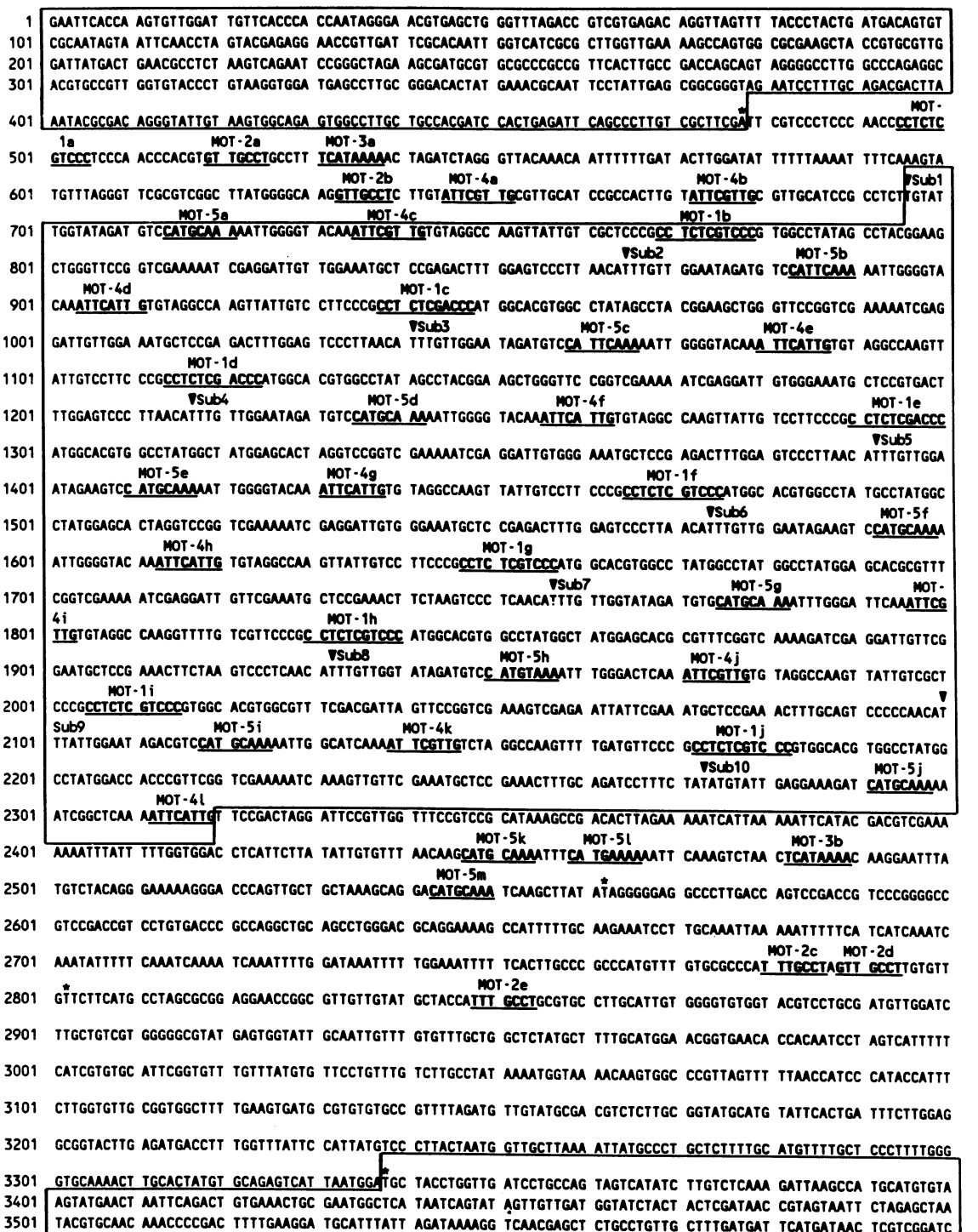

Figure 3. Nucleotide sequence EB3890. The boxed regions from positions 1 through 478, and 3338 through 3876 represent the 25S and 18S rRNA 3′ and 5′ ends, respectively. The subrepeat region—beginning at position 696 and extending through position 2319—contains nine full length subrepeats, one truncated subrepeat, and is also boxed. Short repeated sequences located within the IGS are bold-face-underlined, and numbered according to the order in which they appear. The consensus sequence for each motif is as follows: MOT-1 = CCTCTCGTCCC; MOT-2 = $^{G}/_{T}$TTGCCT; MOT-3 = TCATAAAA; MOT-4 = ATTC$^{G}/_{A}$TTG; MOT-5 = CATGCAAA. The asteric at positions 478, 2562, 2802, and 3338 represents the locations of S1 nuclease generated products.

goodness-of-fit tests. This strongly supports the assignment of the Alaska S variant (corresponding to band b) to the *Rrn1* locus on Chromosome 4 and of the Alaska L variant (corresponding to bands c and f) to the *Rrn2* locus on Chromosome 7. Further support for the assignment of the S variant to Chromosome 4 derives from a tight linkage with *Fa*, mapping an estimated 2 ± 9

map units from the marker (Table 1). This is very close to the estimated distance between *Rrn1* and *Fa* made by Polans et al. (31) using a different cross.

Identification of a short rDNA length variant clone

The two major rDNA length variants in cv. Alaska can be differentiated via restriction enzyme sites located within their IGS. Putative short rDNA length variant clones were identified by the lack of *Eco*RI sites in the subrepeat region (figure 1). The clone, *RRN*pss1, shows an RFLP pattern characteristic of the Alaska short length variant and was subcloned into the DNA vector plasmid pGEM-3 (Promega Biotec; Madison, WI). To determine if pRRNpss1 was a genomic clone representative of the short rDNA length family, we performed two types of restriction site analysis. In the first analysis, genomic DNA and pRRNpss1 DNA was either digested with *Hin*dIII alone, or double-digested with *Hin*dIII plus either *Eco*RI, *Bam*HI or *Sph*I, fractionated on agarose gels, blotted, and probed with the full-length pRRNpss1 clone. The results are shown in figure 2b. Restriction digestion with *Hin*dIII indicates that the pRRNpss1 insert (lanes 2–4) comigrates with the genomic 8.6 kb short rDNA length variant (lane 1). Restriction digestion of pRRNpss1 with *Eco*RI (lane 6) indicates the lack of 1.2 kb and 1.4 kb fragments which are characteristic of the long length variant (lane 5). In addition, the 2.3 kb fragment characteristic of the short variant (lane 6) comigrates with the identical size fragment of genomic rDNA (lane 5). Restriction digestions of pRRNpss1 with *Bam*HI (lane 8) and *Sph*I (lane 10) yield fragments which are identical in size to those resulting from similarly digested genomic DNA (lanes 7 and 9, respectively). The extra band present in the *Bam*HI and *Sph*I digested genomic DNA is derived from the L variant.

Because the short rDNA length variant lacks the *Eco*RI sites present in the IGS of the long length variant, it is possible to use the indirect end-labeling probe EBg541 (figure 1) to examine restriction sites in genomic copies of the short length rDNA variant that lie between the 3′ end of the subrepeat array and the *Eco*RI site in the 18S coding region. Using this method we have confirmed that the following sites, present within the IGS of the short rDNA clone, are also present within the bulk of the genomic short rDNA (data not shown): the 5′-most *Sph*I site at position 1774; the 5′-most *Xmn*I site at position 2462; the *Hin*dIII site at position 2553; the *Xba*I site at position 3490; the *Hinc*II site at position 3550; the *Bam*HI site at position 3871.

Copy number reconstruction experiments (figure 2, lanes 1–4) were performed to determine the number of long length and short length rDNA copies in genomic DNA of cv. Alaska. It has been previously determined that the total number of rDNA copies per haploid genome in pea (cv. Alaska) is approximately 4000 (47). Using a standard amount of genomic DNA and varying amounts of pRRNpss1 DNA, we have determined that the genome of pea (cv. Alaska) contains approximately 1000 copies of the short rDNA variant and 3000 copies of the long length variant.

Nucleotide sequence of pEB3890

DNA fragments were isolated from pEB3890 following digestion with one or a combination of restriction endonucleases (*Eco*RI, *Sph*I, *Xmn*I, *Hin*dIII, or *Bam*HI), cloned into the appropriate sites in M13mp18 or M13mp19, and sequenced. Regions of pEB3890 lacking convenient cloning sites were sequenced from longer fragments by priming with 17 bp synthetic oligonucleotides. The 1.7 kb subrepeat region was sequenced in the following manner: A 1.7 kb *Hin*dIII/*Rsa*I fragment from pUC19 (48) was ligated to the *Eco*RI/*Hin*dIII 2550 bp fragment from pEB3890. Partial digestion of insert from the resulting clone with *Rsa*I or *Nco*I allowed the separation, isolation and subsequent cloning of the partially digested subrepeat fragments.

The complete nucleotide sequence of pEB3890 (3876 bp), shown in figure 3, contains 478 bp of the 25S rRNA 3′ end, 539 bp of the 18S rRNA 5′ end, and 2859 bp of intergenic spacer.

S1 nuclease mapping

Four major S1 nuclease generated products have been detected using probes derived from the subclone EB3890. These products, which map to positions 478, 2562, 2802, and 3338 (denoted by asterics in figure 3), probably represent the 3′ terminus of 25S rRNA (figure 4a), the site of transcript initiation (figure 4b), a processing site (figure 4c), and the 5′ terminus of 18S rRNA (figure 4d), respectively.

The approximate locations of 18S and 25S rRNA in the pea rDNA clone pHA1 have been previously determined using R-loop analysis (35). Based on data from that study, the 25S 3′ border was placed between *Eco*RI and *Bgl*II restriction sites, and the 18S 5′ border was placed between two *Xmn*I restriction sites (see figure 1 of 35). We have used DNA fragments from these two regions as probes in S1 mapping experiments, and observe two S1 generated products at positions 478 and 3358 (figure 3) which represent the 25S rRNA 3′ border and 18S rRNA 5′ border, respectively. A comparison of the pea IGS sequence with published rRNA border sequences of radish (17), cucumber (49), maize (14,15), and wheat (16) reveals rRNA/IGS borders which correlate with our S1 nuclease data.

Two prominant S1 nuclease generated products are observed within the IGS of pea. The upstream-most of these products (figure 4b) maps to position 2562 (figure 3). Table 2 compares the sequence surrounding this S1 nuclease generated product with sequences surrounding the putative sites of initiation in other plant systems, as well as rDNA initiation sites in animal systems. Although rDNA promoters are species-specific, and generally show limited sequence similarity, homology is observed between the sequence surrounding position 2562 in pea and putative sites of initiation in monocots and other dicots. For example, between position −6 and +10 (Table 2), 15/16 residues are identical between the pea and radish sequences. The unique stretch of sequence homology, in conjunction with the fact that no other 5′ transcripts are detected upstream of position 2562 (using both, S1 analysis and primer extension methods) suggests that the S1 nuclease generated product mapping to position 2562 probably represents the primary site of transcript initiation.

The second S1 nuclease generated product located within the IGS (figure 4c) maps to position 2802 (figure 3). Since the sequence surrounding this product shows no homology to sequences surrounding the the pea site of initiation, or published transcription start site sequences from other systems, it probably does not represent a second site of transcript initiation. S1 nuclease generated products are also observed downstream of the transcription start site in radish (17), maize (14,15) and wheat (16). The authors of the above studies propose that the downstream S1 nuclease generated products represent major processing sites. It is probable that the S1 nuclease generated product, which maps to position 2802 in pea, also represents a processing site.

Structural analysis of the intergenic spacer

Located within the intergenic spacer of pea are multiple copies of five short sequences, or motifs, which range in size from 7

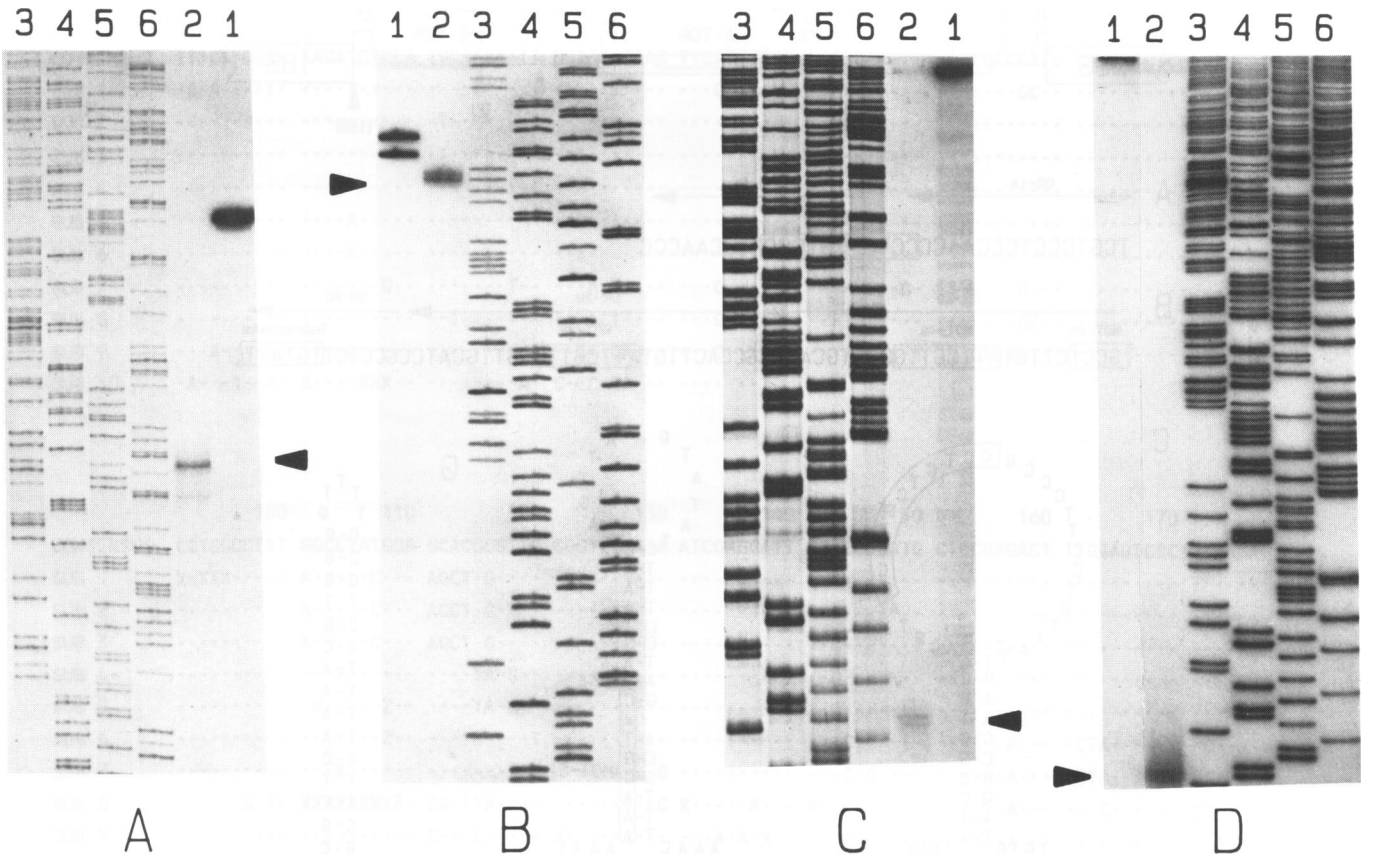

Figure 4. S1 nuclease mapping. (A) The probe HpBg311 (lane 1) generated a 247 bp S1 protected fragment (lane 2) mapping to position 478 (figure 3). (B) The probe HX194 (lane 1) generated a 186 bp S1 protected fragment (lane 2) mapping to position 2562 (figure 3). (C) The probe PR249 (lane 1) generated a 79 bp S1 protected fragment (lane 2) mapping to position 2802 (figure 3). (D) The probe XX663 (lane 1) generated a 73 bp S1 protected fragment (lane 2) mapping to position 3338 (figure 3). M13mp18 sequencing reactions were used as standards in lanes 3 through 6 (G, A, T, and C reactions, respectively).

Table 2. Comparison of RNA polymerase I initiation sites and flanking regions.

+1 T C A A G C T T A T A - T A G G G G G - A G G C C	pea
+1 T A A G T G T T A T A - T A G G G G G T A G G C A	radish (17)
C C T C A G G T A T A G T A G G G G G T A G G G A	maize (14)
C C T C G G G T A T A G T A G G G A G G A G G G G	wheat (16)
C G G G C A G G A A G G T A G G G G A A G A C C G	*Xenopus* (57)
C A A A A A C T ACT A - T A G G T A G G C A G T G	*Drosophila* (58)
G C C G G G T T A T A - T G C T G A C A C G C T G	human (59)

bp to 11 bp; they are numbered 1 through 5 in the order of which they appear within the IGS, and multiple repititions of the same sequence are designated a, b, c, ... (figure 3). Because of their location, reiteration, and what is known about rDNA transcription and regulation in *Xenopus*, *Drosophila*, and the mammals, it is possible that they may be involved with initiation of transcription (motif-5), and termination and/or processing (motif-1 through motif-4). Also located within the IGS are two sets of direct repeats (other than the subrepeats), four stem-loop structures and a palindrome. These structures (figure 5), along with motifs-1 through motif-5, will be discussed in the following sections.

The 25S border. There are two sets of direct repeats near the 25S border (figure 5). Direct repeat-1 (DR-1) begins 2 bp downstream of the 25S rRNA 3′ border, and is comprised of two 16 bp repeats (DR-1a and DR-1b, figure 5a). The first 11 bp of DR-1 shows homology to the 5′ IGS border sequence in radish (17), cucumber (49), maize (14,15), and wheat (16); in these systems, however, the sequence is not part of a direct repeat, as it is in pea. Direct repeat-2 (DR-2) begins 60 bp upstream of the subrepeat array, and is comprised of two full length, and one partial copy of a 27 bp long repeat (DR-2a, DR-2b, and DR-2c, figure 5b). The partial copy (DR-2c) contains only the first 12 bp of DR-2. The 1.7 kb subrepeat region begins at the seventh position of DR-2c. Interestingly, a 15 bp sequence which shows homology to the terminal 15 bases of DR-2 (11/15 match), is located immediately downstream of the subrepeat region (figure 3, positions 2335−2349). Thus, the 1.7 kb subrepeat region appears to be 'inserted' into DR-2c.

Motif-1 (CCTCTCG$^{T}/_{A}$CCC) traverses DR-1a and DR-1b,

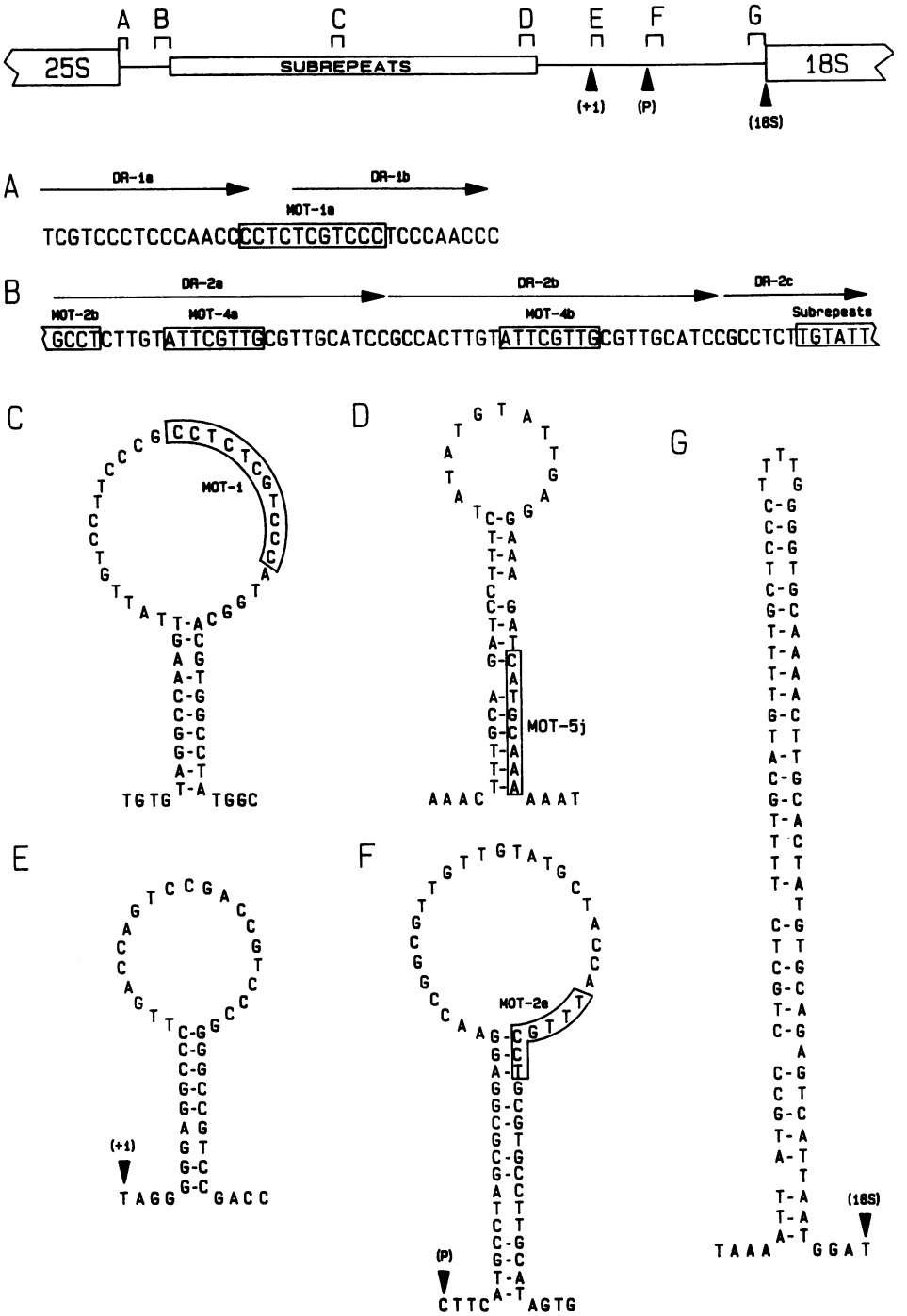

Figure 5. Direct repeats and secondary structures located within the IGS. (A) DR-1, composed of two 16 bp repeats (DR-1a and DR-1b) begins 2 bp downstream of the IGS 5′ border. (B) DR-2, composed of two complete 27 bp repeats (DR-2a and DR-2b) and a 12 bp partial repeat (DR-2c), begins 60 bp upstream of the 1.7 kb subrepeat array. (C) A 10 bp stem-30 base loop structure with a free energy of −14.1 kcal/mol begins 50 bp downstream of the beginning of subrepeats one through nine (figure 6). (D) A 16 bp stem-13 base loop structure with a free energy of −15.0 kcal/mol is located traversing the junction of subrepeats 9 and 10. (E) A 9 bp stem-21 base loop structure with a free energy of −18.0 kcal/mol begins 4 bp downstream of the site of initiation. (F) A 16 bp stem-28 base loop structure with a free energy of −16.6 kcal/mol begins 4 bp downstream of the major processing site. (G) A 74 bp palindrome with a free energy of −38.5 kcal/mol begins 77 bp upstream of the 18S rRNA 5′ border. Arrowheads in E, F and G represents the site of initiation (+1) and processing (P), and the start of the 18S rRNA coding region (18S), respectively.

and occurs once within each of the nine full-length subrepeats. Motif-2 ($^{G}/_{T}$TTGCCT) occurs 41 bp and 155 bp downstream of the IGS 5′ border. Motif-2 is also located 22 bp and 14 bp upstream of, and 46 bp downstream of the putative processing site (figure 3). In wheat (16), five copies of motif-2 are located upstream of the major processing site. The locations of motif-2 in both pea and wheat are consistant with a role in processing.

The subrepeat region. A 1.7 kb subrepeat region begins 218 bp downstream of the IGS 5′ border. This region is comprised of

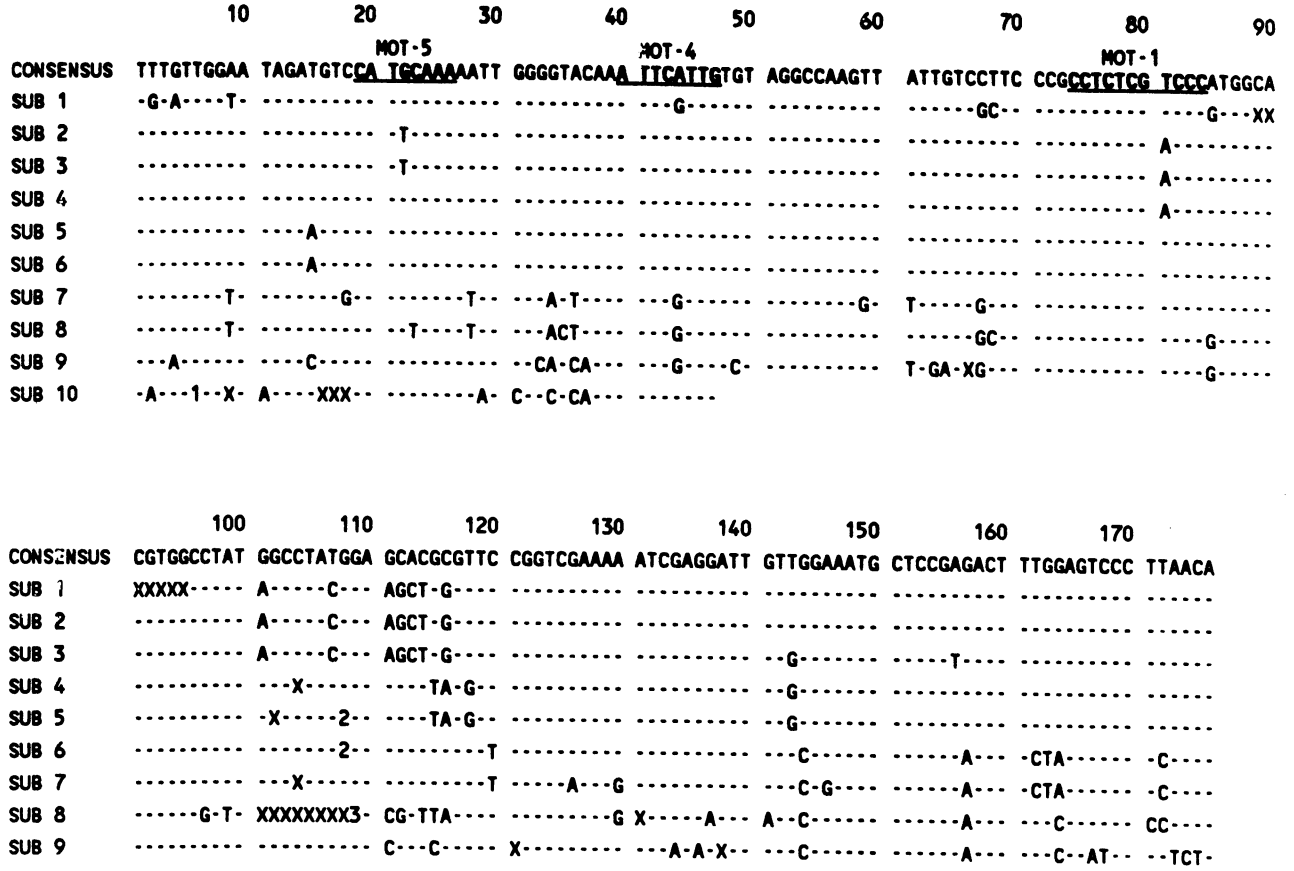

Figure 6. Subrepeat sequences. The subrepeats were manually aligned to achieve maximum homology. A consensus sequence was determined and is displayed in the top row. Nucleotides within the subrepeats which are identical to those in the consensus sequence are denoted (−), while those which are present in the consensus but not in the subrepeats are denoted (X). Insertion of the sequences TATTGA in subrepeat 10, GCCTATGG in subrepeats 5 and 6, and CG in subrepeat 8 are denoted (1), (2), and (3), respectively.

Table 3. Comparison of length, thermodynamic stability, and location of palindromic sequences near the 18S rRNA 5′ border.

ORGANISM	PALINDROME LENGTH	STABILITY	LOCATION
Pea	74 bp	−38.5 kcal/mol	−77 bp
Radish (17)	90 bp	−32.2 kcal/mol	−90 bp
Cucumber (49)	83 bp	−29.9 kcal/mol	−110 bp
Wheat (16)	65 bp	−19.7 kcal/mol	−73 bp
Maize (14)	58 bp	−24.5 kcal/mol	−68 bp
Drosophila (60)	60 bp	−9.3 kcal/mol	−60 bp
Sea urchin (61)	51 bp	−24.8 kcal/mol	−52 bp

nine full length subrepeats, which range in size between 168 bp and 183 bp, and a single truncated subrepeat (located at the 3′ end of the array), which is 47 bp long. Subrepeats 2, 3, 4, 5 and 6 show 95%, 94%, 97%, 93% and 91% homology to a subrepeat consensus sequence (figure 6), respectively, while subrepeats 1, 7, 8, 9 and 10 show less homology (88%, 89%, 80%, 84% and 64%, respectively). The pattern of greater divergence from the consensus sequence among the external subrepeats is also common to the rDNA subrepeat array in maize (14) and wheat (16).

Computer analysis of the subrepeat consensus sequence (figure 6) indicates the possible formation of two stem-loop structures. The first is a 10 bp stem-30 base loop structure (figure 5c), with a free energy of −14.1 kcal/mol; it can form in each of the first nine subrepeats. The loop region of this secondary structure contains a complete copy of motif-1. The only other copy of motif-1 is located at the 25S rRNA border where it traverses DR-1a and DR-1b (figure 5a). Direct repeats and loop domains of stem-loop structures may be effective ways of ensuring that if transcribed, motif-1 occurs in a single stranded form. The second structure, containing a 16 bp stem and 13 base loop (figure 5d), has a free energy of −15.0 kcal/mol. Unlike the stem-loop structure described above which occurs within each subrepeat, this structure occurs only once within the subrepeat array traversing the border between subrepeats 9 and 10. The stem region of this structure contains a complete copy of motif-5,

Table 4. Periodicity of guanine residues (G) on the noncoding strand which lie on the same face of the DNA helix.

−20	−10	+1	
. . . . . *. . . . . *. . . . . *.			
G	G	G	G Mouse (53)
G	G	G	G Pea
G	G	G	G Radish (17)
GG	G	G	G Maize (14)
G		G	G Wheat (16)
G	G	G	G Human (59)
G	G		G *Drosophila* (58)
	G	G	G *Xenopus* (57)
	G	G	G Yeast (62)
	G	G	G *Neurospora* (63)
G	G	G	G *Tetrahymena* (64)
G	G	G	G *E. coli* (65)
GG	G	G	G *Acanthamoeba* (12)

Table 5. Comparison of stem size, loop size, thermodynamic stability, and location of secondary structures near the rDNA site of initiation.

ORGANISM	STEM	LOOP	STABILITY	LOCATION
Pea	9 bp	21 b	−18.00 kcal/mol	+4
Radish (17)	10 bp	6 b	−11.30 kcal/mol	+2
Wheat (16)	7 bp	6 b	−9.50 kcal/mol	+1
Maize (14)	6 bp	5 b	−14.50 kcal/mol	−26
Xenopus (57)	7 bp	14 b	−13.54 kcal/mol	−13
Drosophila (58)	8 bp	7 b	−13.60 kcal/mol	+1
Sea urchin (61)	11 bp	11 b	−19.94 kcal/mol	−18
Human (59)	7 bp	8 b	−8.70 kcal/mol	+14
Mouse (66)	8 bp	13 b	−8.41 kcal/mol	+11
Rat (67)	7 bp	16 b	−6.14 kcal/mol	+12

which we believe to be a core promoter/enhancer element (see following section). If transcription of the subrepeat region occurs in pea, as it does in other systems (reviewed in 25), then this copy of motif-5 would occur in a double stranded RNA form.

The transcript initiation region. S1 analysis (figure 4c) and primer extension data (not shown) have mapped the site of transcript initiation to position 2562 (figure 3). Motif-5 (CATGCAAA) occurs 19 bp upstream of this site of initiation, two more times at −115 and −103 (relative to the site of initiation) and once within each of the 10 subrepeats. This motif represents the only sequence which is reiterated both in the promoter region, and in the subrepeats. In *Xenopus* (7,21,50,51), *Drosophila* (22), rat (8), mouse (19) and human (4), sequences located upstream of, and showing homology to the rDNA promoter have been demonstrated to act as enhancer elements, and in some cases, initiate transcription. By anology to these systems, it is very likely that motif-5 represents a core promoter/enhancer element in pea. We are currently testing this idea using the mobility shift and transient expression assays.

Motif-3 (TCATAAAA) is located 53 bp downstream of the 25S rRNA 3′ border, and again 81 bp upstream of the site of initiation. In *Xenopus* (52) and mouse (27,28), a short sequence (7 bp and 18 bp, respectively)−located downstream of the 28S rRNA 3′ border (235 bp and 570 bp, respectively)−directs termination of the primary transcript; an identical copy of the terminator sequence−located 215 bp (*Xenopus*) and 171 bp (mouse) upstream from the site of initiation−functions as a failsafe terminator and probably prevents promoter occlusion (29). A similar situation exists in human (4) where an 11 bp sequence−located 174 bp upstream of the site of initiation and showing homology to the mouse terminator−is essential for elevated promoter activity. Based on anology to these systems, it is possible that the copy of motif-3 located 53 bp downstream of the 25S rRNA 3′ border is involved with termination, while the copy located 81 bp upstream of the site of initiation is involved with failsafe termination.

A 9 bp stem-21 base loop structure with a free energy of −18.0 kcal/mol begins 4 bp downstream of the site of initiation (figure 5e). Similar structures are identified near the site of initiation in other plant and animal systems (Table 3). Because of this conservation, it is possible that these structures may be involved with the regulation of rDNA transcription.

The region surrounding the site of initiation in pea contains G residues at positions −43, −34, −25, −16, −7, +3 and +10 −all lying on the same face of the helix. Analysis of published rDNA promoter sequences of plants and animals, as well as of *E. coli*, shows G residues at positions analogous to those observed in pea (Table 4). In mouse, a G to A base change at position −25, −16, or −7 greatly reduces or abolishes transcription (53,54). The fact that all rDNA systems examined contain G residues on the same face of the rDNA helix indicates a conserved G residue periodicity which may be necessary for transcription factor binding.

The promoter region of the short rDNA length variant contains a developmentally regulated *Hpa*II methylation site (36) located at +31, and three developmentally-regulated DNAseI hypersensitive sites (38) located at −100, +150 and +350. In wheat (16), a *Hpa*II restriction site is located 141 bp downstream of the site of initiation; methylation of this site is under genetic control (24). Radish (17) and maize (14,15) contain *Hpa*II sites located 105 bp and 30 bp downstream from the site of initiation,

respectively, but methylation studies have not yet been carried out using these systems.

The processing region. One major S1 nuclease generated product is observed 240 bp downstream from the site of initiation, at position 2802 (figure 3). This product is analogous to S1 generated products located downstream of the site of initiation in radish (17), maize (14,15) and wheat (16), and probably represents a primary processing site (see 'S1 nuclease mapping' section). A computer analysis of sequence surrounding this processing site reveals the presence of a 16 bp stem-28 base loop structure, with a free energy of −16.6 kcal/mol (figure 5f), beginning 4 bp downstream of the processing site.

The 18S border. A 74 bp palindrome with a free energy of −38.5 kcal/mol is found in the rDNA of pea beginning 77 bp upstream of the 18S rRNA 5′ border (figure 5g). An analysis of published sequence data from other plant and animal systems indicates the presence of a thermodynamically stable palindrome in the region analogous to that of pea (Table 5). The fact that large sequences containing dyad symmetry are present immediately upstream of the 18S border in both plants and animals strongly suggests the involvement of palindromes in the processing of 18S RNA. It is interesting to note that in yeast (55) and rice (56), a palindrome with a free energy of −18.4 and −19.3 kcal/mol, respectively, is located in the internal transcribed spacer at the 25S rRNA 5′ border, suggesting that palindromes may also function in processing at this location.

ACKNOWLEDGMENTS

The authors would like to thank Kevin Folta for technical assistance, and Drs. M. Erhart, L. Miller, and B. Nichols for critical reading of the manuscript. KJP gratefully acknowledges an LCMDB research fellowship. This work was supported by USDA grant 88-37261-3734 to LSK, and Biomedical Research Support Grant BRSG-S07-RR07176 and Northern Illinois University Graduate School Fund to NP.

REFERENCES

1. Reeder, R. H. (1984). Cell, **38**, 349−351.
2. Rogers, S. O. & Bendich, A. J. (1987). Plant Mol. Biol. **9**, 509−520.
3. Sollner-Webb, B. & Tower, J. (1986). Ann. Rev. Biochem. **55**, 801−830.
4. Haltiner-Jones, M., Learned, R. M. & Tjian, R. (1988). Proc. Natl. Acad. Sci., U.S.A. **85**, 669−673.
5. Learned, R. M., Smale, S. T., Haltiner, M. & Tjian, R. (1983). Proc. Natl. Acad. Sci., U.S.A. **80**, 3558−3562.
6. Grummt, I. (1981). Nucl. Acids Res. **9**, 6093−6102.
7. Miller, K., Tower, J. & Sollner-Webb, B. (1985). Mol. Cell Biol. **5**, 554−562.
8. Cassidy, B. G., Yang-Yen, H. & Rothblum, L. I. (1986). Mol. Cell Biol. **6**, 2766−2773.
9. Moss, T. (1982). Cell, **30**, 835−842.
10. Sollner-Webb, B., Wilkinson, J. K., Roan, J. & Reeder, R. H. (1983). Cell **35**, 199−206.
11. Kohorn, B. & Rae, P. M. M. (1983). Proc. Natl. Acad. Sci., U.S.A. **50**, 3263−3268.
12. Iida, C. T., Kownin, P., & Paule, M. R. (1985). Proc. Natl. Acad. Sci., U.S.A. **82**, 1668−1672.
13. Kownin, P., Iida, C. T., Brown-Shimer, S. & Paule, M. R. (1985). Nucl. Acids Res. **13**, 6237−6248.
14. McMullen, M. D., Hunter, B., Phillips, R. L. & Rubenstein, I. (1986). Nucl. Acids Res. **14**, 4953−4968.
15. Toloczyki, C. & Feix, G. (1986). Nucl. Acids Res. **14**, 4969−4986.
16. Barker, R. F., Harberd, N. P., Jarvis, M. G. & Flavell, R. B. (1988). J. Mol. Biol. **201**, 1−17.
17. Delcasso-Tremousaygue, D., Grellet, F, Panabieres, F, Ananier, E. D. & Delseny, M. (1988). Eur. J. Biochem. **172**, 767−776.
18. Elion, E. A. & Warner, J. R. (1986). Mol. Cell Biol. **6**, 2089−2097.
19. Kuhn, A. & Grummt, I. (1987). EMBO J. **6**, 3487−3492.
20. Cassidy, B. G., Yang-Yen, H. & Rothblum, L. I. (1987). Mol. Cell Biol. **7**, 2388−2396.
21. Labhart, P. & Reeder, R. H. (1984). Cell **37**, 285−289.
22. Coen, E. S. & Dover, G. A. (1982). Nucl. Acids Res. **10**, 7017−7026.
23. Grimaldi, G. & DiNocera, P. P. (1986). Nucl. Acids Res. **14**, 6417−6432.
24. Thompson, W. F., Flavell, R. B., Watson, J. C. & Kaufman, L. S. (1988). In The architecture of eukaryotic genes (Kahl, G., ed), pp. 386−396, VCH Verlagsgesellschaft, Frankfurt, Germany.
25. Baker, S. M. & Platt, T. (1986). Cell **47**, 839−840.
26. McStay, B. & Reeder, R. H. (1986). Cell **47**, 913−920.
27. Grummt, I., Kuhn, A., Bartsch, I. & Rosenbauer, H. (1986). Cell **47**, 901−911.
28. Henderson, S. & Sollner-Webb, B. (1986). Cell, **47** 891−900.
29. Bateman, E. & Paule, M. R. (1988). Cell **54**, 985−992.
30. Ellis, T. H. N., Davis, D. R., Castelton, J. A. & Bedford, I. D. (1984). Chromosoma **91**, 74−81.
31. Polans, N. O., Weeden, N. F. & Thompson, W. F. (1986). Theor. Appl. Genet. **72**, 289−295.
32. Folkeson, D. (1984). Heriditas **101**, 227−233.
33. Lamm, R. (1951). Heriditas **37**, 356−372.
34. Lamm, R. (1981). Heriditas **94**, 45−52.
35. Jorgensen, R. A., Cuellar, R. E., Thompson, W. F. & Kavanagh, T. A. (1987). Plant Mol. Biol. **8**, 3−12.
36. Watson, J. C., Kaufman, L. S. & Thompson, W. F. (1987). J. Mol. Biol. **193**, 15−26.
37. Baerson, S. R. & Kaufman, L. S. (1990). Mol. Cell Biol. **10**, 842−845.
38. Kaufman, L. S., Watson, J. C. & Thompson, W. F. (1987). Proc. Natl. Acad. Sci., U.S.A. **84**, 1550−1554.
39. Polans, N. O., Weeden, H. F. & Thompson, W. F. (1985). Proc. Natl. Acad. Sci., U.S.A. **82**, 5083−5087.
40. Loenen, W. A. M. & Brammer, W. J. (1980). Gene **20**, 249−259.
41. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982). In Molecular cloning: a laboratory manual. (Cold Spring Harbor), Cold Spring Harbor, N. Y.
42. Messing, J. (1983). Meth. Enzym. **101**, 20−78.
43. Barr, P. J., Thayer, R. M., Laybourn, P., Najarian, R. C., Seela, F. & Tolan, D. R. (1986). BioTechniques **4**, 428−432.
44. Lipman, D. J. & Pearson, W. R. (1985). Science **227**, 1435−1441.
45. Murray, M. G. & Thompson, W. F. (1980). Nucl. Acids Res. **8**, 4321−4325.
46. Suiter, K. A., Wendel, J. F. & Case, J. S. (1983). J. Hered. **74**, 203−204.
47. Ingle, J., Timmins, J. N. & Sinclair, J. (1975). Plant Phys. **55**, 496−501.
48. Yanish-Perron, C., Vieira, J. & Messing, J. (1985). Gene **33**, 103−119.
49. Ganal, M., Torres, R. & Hemleben, V. (1988). Mol. Gen. Genet. **212**, 548−554.
50. Pikaard, C. S. & Reeder, R. H. (1988). Mol. Cell. Biol. **8**, 4282−4288.
51. Reeder, R. H., Roan, J. G. & Dunaway, M. (1983). Cell **35**, 449−456.
52. Labhart, P. & Reeder, R. H. (1986). Cell **45**, 431−443.
53. Kishimoto, T., Nagamine, M., Sasaki, T., Takakusa, N., Miwa, T., Kominami, R. & Muramatsu, M. (1985). Nucl. Acids Res. **13**, 3515−3552.
54. Skinner, J. A., Ohrlein, A. & Grummt, I. (1984). Proc. Natl. Acad. Sci., U.S.A. **81**, 2137−2141.
55. Schaak, J., Mao, J. & Soll, D. (1982). Nucl. Acids Res. **10**, 2851−2864.
56. Takaiwa, F., Oono, K. & Sugiura, M. (1985). Plant Mol. Biol. **4**, 355−364.
57. Sollner-Webb, B. & Reeder, R. H. (1979). Cell **18**, 485−499.
58. Long, E. O., Reebert, M. L. & Dawid, I. B. (1981). Proc. Natl. Acad. Sci., U.S.A. **78**, 1513−1517.
59. Financsek, I., Mizumoto, K., Mishima, Y. & Muramatsu, M. (1982). Proc. Natl. Acad. Sci., U.S.A. **79**, 3092−3096.
60. Simeone, A., LaVolpe, A. & Boncinelli, E. (1985). Nucl. Acids Res. **13**, 1089−1101.
61. Ho, K. & Stafford, D. W. (1985). Gene **39**, 49−54.
62. Balzi, E., DiPietro, A., Goffeau, A., vanHeerikhuizen, H. & Klootwijk, J. (1985). Gene **39**, 165−172.
63. Tyler, B. M. & Giles, N. H. (1985). Nucl. Acids Res. **12**, 4311−4332.
64. Saiga, H., Mizumoto, K., Matsui, T. & Higoshinakagawa, T. (1982). Nucl. Acids Res. **10**, 4223−4236.
65. Lindahl, L. & Zengel, J. M. (1986). Ann. Rev. Genet. **20**, 297−326.
66. Urano, Y., Kominami, R., Mishima, Y. & Muramatsu, M. (1980). Nucl. Acids Res. **8**, 6043−6058.
67. Financsek, I., Mizumoto, K. & Muramatsu, M. (1982). Gene **18**, 115−122.