J Mol Evol (1993) 36:369–379

Journal of Molecular Evolution

Fine Structure and Evolution of the rDNA Intergenic Spacer in Rice and Other Cereals

F. Cordesse, R. Cooke, D. Tremousaygue, F. Grellet, and M. Delseny

Laboratoire de Physiologie et Biologie Moléculaire Végétales, URA 565 du Centre National de la Recherche Scientifique, Université de Perpignan, 66860 Perpignan-Cedex, France

Summary. The intergenic spacer of a rice ribosomal RNA gene repeating unit has been completely sequenced. The spacer contains three imperfect, direct repeated regions of 264–253 bp, followed by a related but more highly divergent region. Detailed analysis of the sequence allows the presentation of an evolutionary scenario in which the 264–253-bp repeats are derived from an ancestral 150-bp sequence by deletion and amplification. Comparison of the rice sequence with those of maize, wheat, and rye shows that, despite considerable divergence from the ancestral sequence, several regions have been highly conserved, suggesting that they may play an important role in the structure and/or expression of the ribosomal genes.

Key words: *Oryza sativa* — Ribosomal DNA — Ribosomal RNA — Intergenic spacer — Repeated sequence

Ribosomal repeat units in plants and animals are arranged in tandem arrays and contain the genes for 25S, 18S, and 5.8S RNAs. Whereas the coding regions for these three RNAs are very highly conserved between even distantly related organisms, the nontranscribed sequences and those which are removed during processing are much less well conserved (Rogers and Bendich 1987; Hemleben et al. 1988). This can be explained by the particular nature of these genes as, in contrast to the considerable number of different genes transcribed by RNA polymerases II and III, only one type of gene, present in multiple copies, is transcribed by one polymerase, thus facilitating coevolution of the genes and the polymerase and other factors implicated in their transcription (Dover and Flavell 1984). The region between the end of the 25S gene and the 18S gene, the intergenic spacer (IGS), is of particular interest as it contains sequences essential for the initiation of transcription, RNA processing signals, the transcription termination site, and, possibly, sequences implicated in the replication of ribosomal DNA (Van't Hof and Lamm 1991). The IGS is characterized by the presence of one or more repeating motifs, which may vary in number and which have been shown to function as enhancers of transcription in *Xenopus*. In most cases the transcription initiation site lies downstream from these repeats, although in some organisms repeats may be present in the transcribed region (Grellet et al. 1989; Barker et al. 1988). In addition, at least in animals the rDNA transcription system is in part species specific (Sollner-Webb and Tower 1986).

Whereas the use of in vitro transcription systems has allowed considerable progress in the understanding of the mechanisms and characterization of the sequences involved in the control of ribosomal gene transcription in animals, much less is known about the control of the expression of these genes in plants. In order to progress in the understanding of basic mechanisms of rDNA transcription it is first

Abbreviations: IGS (ribosomal gene intergenic spacer), rDNA (ribosomal DNA), rRNA (ribosomal RNA)

Offprint requests to: M. Delseny

necessary to obtain sequence information. We have characterized rDNA spacer sequences of radish (Delcasso-Tremousaygue et al. 1988) and *Brassica oleracea* (Tremousaygue et al. 1992) and have begun the characterization of specific proteins binding to the radish rDNA spacer (Echeverria et al. 1992) as well as to related crucifers. Several IGS sequences have been determined in other dicots (see Zentgraf et al. 1990 and our Fig. 4 for references) as well as for several cereals: maize (Toloczyki and Feix 1986; McMullen et al. 1986), wheat (Lassner and Dvorak 1986; Lassner et al. 1987; Barker et al. 1988), and rye (Appels et al. 1986). We undertook the characterization of an rDNA spacer from rice since it is an important crop.

In a preliminary step toward the characterization of the IGS from rice, we carried out a study on the length heterogeneity of the ribosomal IGS from different accessions (Cordesse et al. 1990) using a clone containing the complete ribosomal repeating unit as a probe (Takaiwa et al. 1984). This demonstrated that rDNA spacers from different rice accessions differ by a regular increment of 260 bp, a result obtained independently by Sano and Sano (1990). In a second step, this region was dissected and various restriction fragments were used to analyze their genome specificity. Cytogenetic studies have revealed several genome types in the Oryzae tribe. All cultivated rice and their presumed ancestors belong to the AA genome group. The other wild species belong to various diploid or allotetraploid types. We demonstrated that the rice rDNA spacer contained 250–260-bp subrepeats which, in AA genome species, cross-hybridize with homologous regions of wild-type species belonging to the BB, CC, and BBCC genomes but not with the EE genome. On the other hand, a short sequence downstream from the subrepeats is highly genome specific and present only in species of the AA genome (Cordesse et al. 1992). These results led us to investigate the structure of the IGS in more detail. As no sequence data were available at the time this study was initiated, we undertook to sequence the IGS from this genome. We show here that this IGS effectively contains three well-conserved ca. 260-bp repeats and an adjacent related sequence. A detailed analysis of this sequence allowed us to trace the evolution of this region of the spacer from an ancestral element.

The IGS sequences of closely related wheat and rye have been compared by Flavell (1989). We have undertaken a comparison of the rice sequence with those of these two species and also that of maize, which is phylogenetically more closely related to rice. The analysis of conserved and variable regions allows us to make more general conclusions about the organization of cereal rDNAs.

Materials and Methods

DNA Isolation, Cloning, and Sequencing. The original plasmid, RR217, containing a complete rice ribosomal repeat unit, was a kind gift from Dr. F. Takaiwa (Takaiwa et al. 1984). All procedures involving manipulations of nucleic acids and cloning were carried out using standard methods described in Maniatis et al. (1982). Restriction enzymes were obtained from Amersham (France), Boehringer Mannheim, BRL, and Appligene. Restriction fragments for sequencing were subcloned into plasmid pUC18 (Yanisch-Perron et al. 1985). Most sequencing was carried out on double-stranded plasmids after alkaline denaturation by the method of Chen and Seeburg (1985), based on that of Sanger et al. (1977) using the pUC sequencing kit from Boehringer Mannheim. Regions which were difficult to subclone for sequencing by this method were recovered as short restriction fragments and sequenced according to Maxam and Gilbert (1977) with a kit from NEN. Reaction products were separated on a 6% polyacrylamide-urea sequencing gel.

Sequence Analysis. Most sequence analysis was carried out using programs designed for or adapted to the Apple Macintosh. Basic analysis was carried out using DNA Strider (Marck 1988). Dot plots were made with Don Gilbert's Dotty Plotter program from the IuBio archive. Matches for 19 nucleotides in a window of 30 residues (corresponding to 63% identity) were plotted. Multiple sequence alignments were carried out with Clustal (Higgins and Sharp 1989) and potential secondary structures were found using the Staden program (Staden 1989) implanted on the Bisance server at the CITI2 in Paris (Bisance, Ministère de la recherche et de la technologie).

Results

Sequence of the Intergenic Spacer

Plasmid RR217 (Takaiwa et al. 1984), which we had previously used in studies on length polymorphism and genome specificity (Cordesse et al. 1990, 1992), contains an EcoRI fragment covering a complete ribosomal repeat unit. The IGS region was subcloned as a ca. 3,000-bp EcoRI-BamHI fragment which was split in smaller restriction fragments subcloned into pUC vectors. The sequence was determined on both strands on most of the fragments and through each restriction site used for subcloning. It differs from the data concerning the same clone, published independently by Takaiwa et al. (1990) at a number of sites and by several deletions. It should be noted that the sequencing of plant ribosomal spacers presents particular problems, as already pointed out (Tremousaygue et al. 1992). First, due to the fact that they contain repetitive motifs, it is important to check very carefully the length of the sequenced fragments against their length measured by gel electrophoresis in order not to miss one or more of these motifs. Second, as they do not code for proteins, sequencing errors cannot be corrected by comparison with the derived amino acid sequence, as is often the case. Our sequencing data

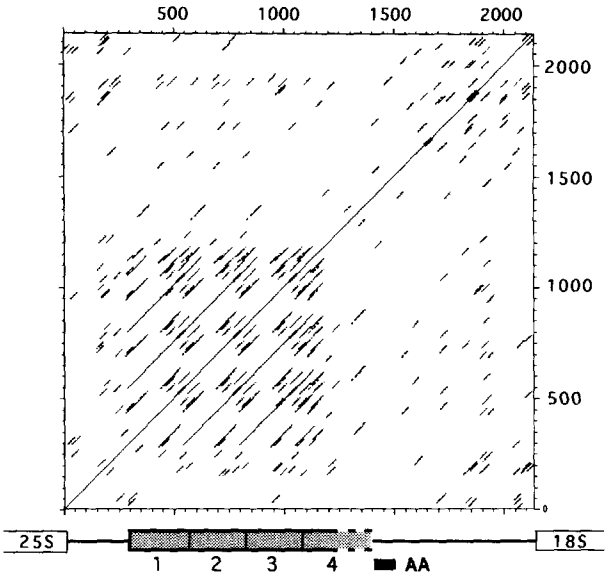


Fig. 1. Dot-matrix plot of the rice intergenic spacer. The sequence of the rice IGS is compared with itself using a window of 30 residues. Each dot represents a match of 19 residues within a window. A schematic representation of the spacer is shown under the plot, indicating the 25S and 18S RNA coding regions, the repeat region, and the region which is specific to the AA genome (AA).

correlate very well with the overall length of the spacer as determined by other methods. All differences with the published sequence were carefully checked and some of the determinations were repeated until unambiguous data were obtained. The discrepancies most likely result from reading mistakes and erroneous interpretation of ambiguous data. Rather than publishing the entire sequence again here, the corrected sequence has been sent to the EMBL databank with a list of the discrepancies. It is available under accession number X58275.

Description of the Repeat Region

The overall structure of the IGS is shown in Fig. 1 as a dot-matrix plot in which the sequence was compared with itself. This presentation most clearly shows the division into repetitive and unique regions. The sequence shows a typical arrangement of a ribosomal IGS, with three direct, imperfect repeats of 264, 253, and 253 bp, respectively, followed by a more highly divergent but related region which might correspond to one or more truncated and partially deleted repeats. Apparently related sequences can also be detected upstream from the first repeat, while immediately downstream is the region specific to the AA genome (Cordesse et al. 1992). However, the dot matrix shows that the repeat structure is much less regular than for other organisms, such as wheat, where the individual repeating units are clearly visible (Barker et al. 1988). In fact, it is possible to detect subrepeats within each of the repetitive elements as shown in Fig. 2a and b. The three direct repeats can be divided into three homologous subregions of about 150 bp (repeats 1-1, 2-1, and 3-1), 60 bp (1-2, 2-2, and 3-2), and 40 bp (1-3, 2-3, and 3-3), the last two types containing increasingly longer deletions compared with the 150-bp sequence. The alignment of these repeats is shown in Fig. 2a and a schematic drawing of the arrangement of the various elements is shown in Fig. 2b. The first part of the first repeat (repeat 1-1) contains an insertion of 11 bp, which is an imperfect repeat of the sequence directly upstream. Four short direct repeats, G(A)TTCGGTC, are present in the 264–253-bp repeat. This short sequence flanks the 150-bp region and is present at the beginning of the 60-bp region as well as at the end of the 40-bp region. The boundary between the last two regions is composed of a short, direct repeat, CCAC, the longer motif having apparently been deleted.

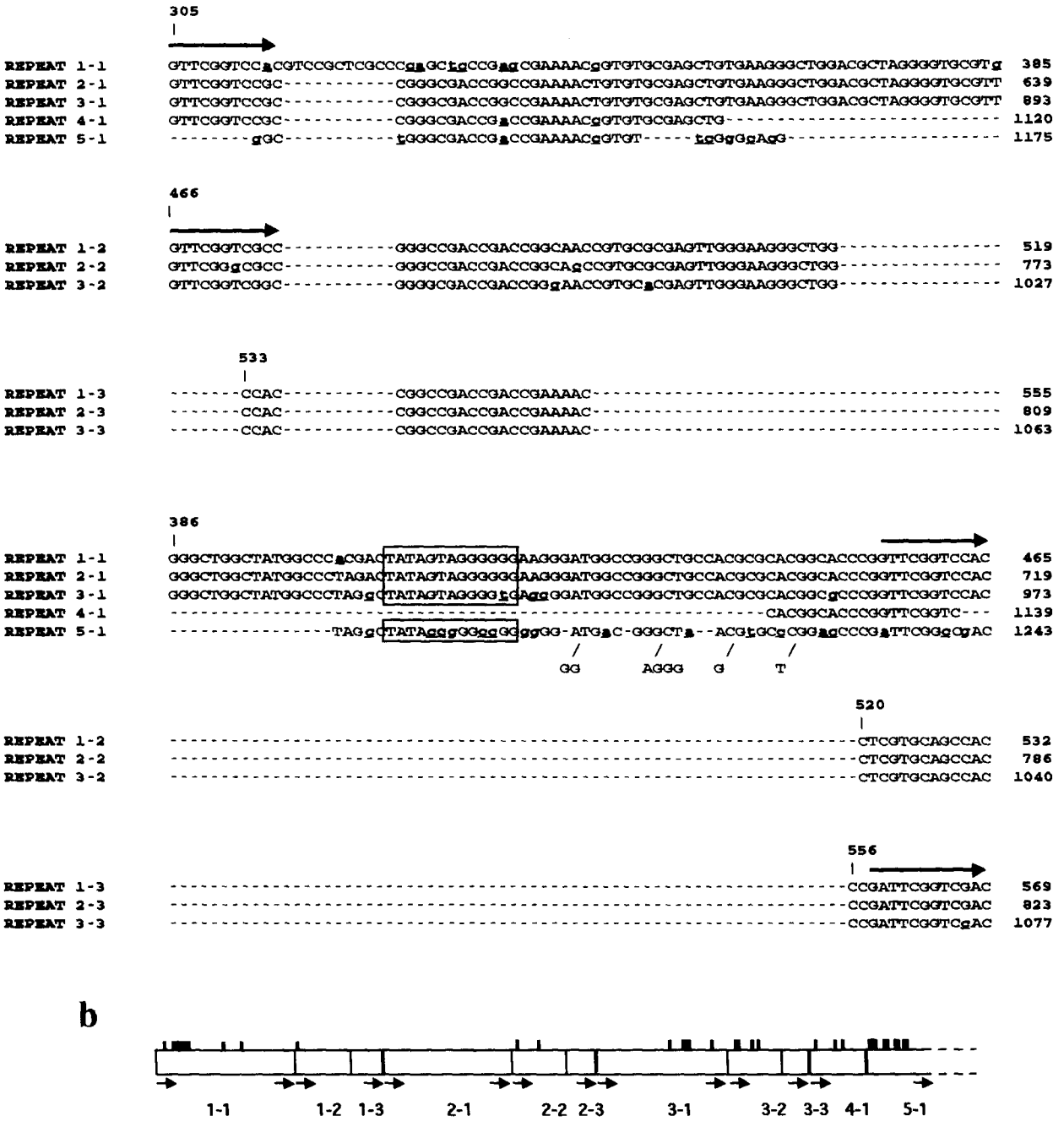
The downstream related region differs from the three repeats in that it has undergone considerably more deletion and mutation. The sequences presented as repeats 4-1 and 5-1 in Fig. 2 are probably derived from highly deleted and mutated versions of at least two 150-bp units, as discussed in the next section. Repeat 4-1 resembles the 150-bp motifs with the central region deleted, ending in the GTTCGGTC sequence. Repeat 5-1 begins with a sequence similar to the 60-bp motifs but contains a GAAAAC sequence which is the same as that in the 150- and 40-bp regions. This is followed by a region which is most similar to those containing the putative transcription initiation sites described below although it has a considerable number of deletions, insertions, and mutations compared with these sequences. The sequence ends with a GATTCGGTC motif as in the 40-bp repeats.

Origin and Evolution of the Repeats

Classical models of repeated DNA sequence evolution, including that of rDNA spacer subrepeats, usually involve successive cycles of amplification and divergence of a short original sequence (Grellet et al. 1986; Dover 1986; Barker et al. 1988; Flavell 1989). However, such a model does not fit well with the observed organization of the rice repeated region and we propose here another scenario involving essentially extensive deletions.

The fact that each 264–253-bp subrepeat is made of three types of related elements suggests that they have all evolved from an ancestral element of 150 bp similar to repeat 2-1 and 3-1. This element has been duplicated several times and various observations suggest that there might be initially as many as

a



b

Fig. 2. Structure of the direct repeats. **a** Alignment of the subregions in each repeat. The coordinates of each sequence on the IGS are shown. Residues which differ from the consensus sequence are shown *underlined* and in *lowercase*. *Dashes* represent absent nucleotides, while the inserted bases in repeat 5-1 (see text) are shown below the sequence. *Arrows* indicate the direct repeats flanking subrepeats. The four potential transcription initiation sites are *boxed*. **b** Schematic representation of the structure of the repeat region. *Arrows* show direct repeats separating subrepeats; *dots* show the localization of residues which differ from the consensus sequence.

6 such units (numbered 1 to 6 in Fig. 3). The observation that the GTTCGGTC sequence is missing at the beginning of repeats 1-3, 2-3, and 3-3, leaving the short direct repeat CCAC (Fig. 2) which is reminiscent of an excision event footprint, suggests that the third 150-bp element has been deleted as well as the beginning of element 4, renamed 4′. Then repeats 2 and 4′ have undergone deletion of their central part, giving rise to a 60-bp 2′ and a 40-bp 4″ element. However, because the two deletions do not overlap perfectly they must have occurred independently. Elements 1, 2′, and 4″ correspond to repeats 1-1, 1-2, and 1-3, respectively, shown in Fig. 2. One possibility is that they result from con-

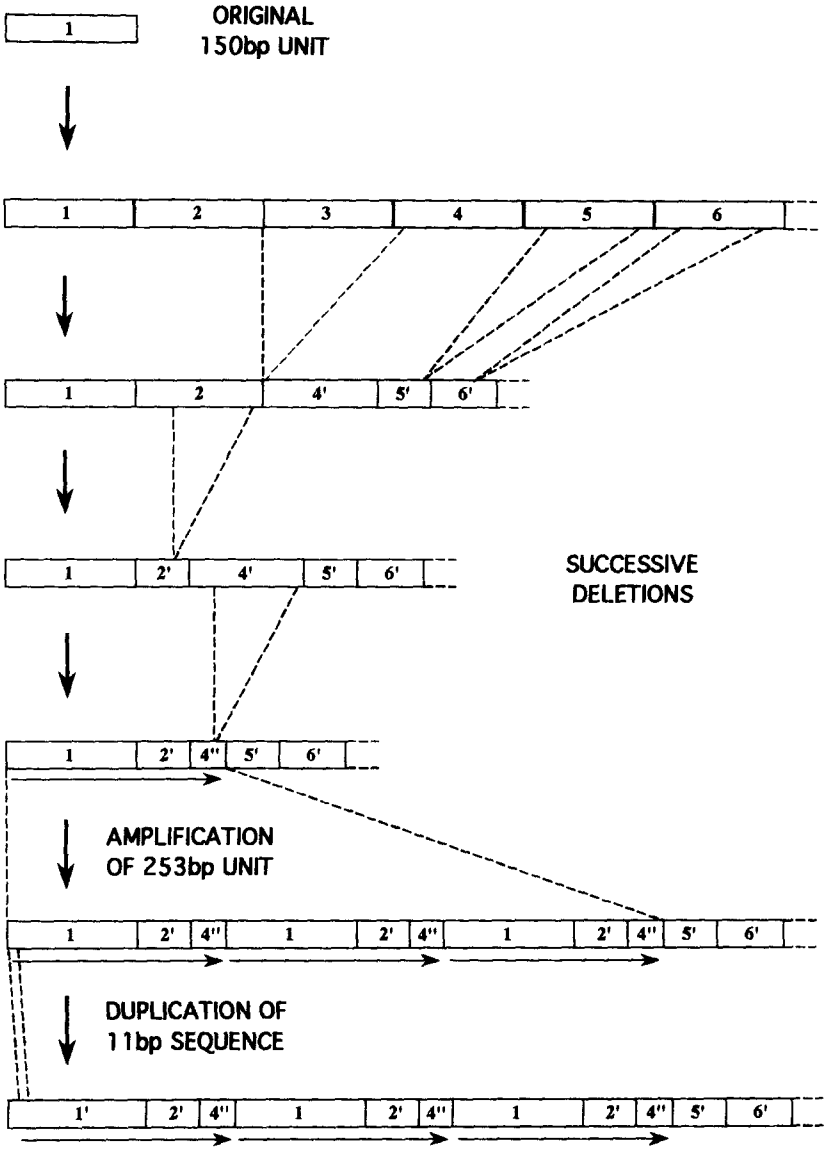


Fig. 3. Origin and evolution of the repeats. The figure shows how the present-day IGS structure may be derived from successive amplification and deletion events, based on an original 150-bp sequence. The chronological order of certain of the events may differ from that shown.

version events between a unit with deletion and another one without. Since we have three almost perfect repeats of 250 bp we have to postulate that a new round of amplification occurred later on. Since the 11-bp insertion is present only in repeat 1-1, this is probably a rather recent event which occurred after the amplification round.

Analysis of the base differences indicated in Fig. 2a and presented schematically in Fig. 2b shows that they are clustered in the first half of repeat 1-1, the second half of repeat 3-1, most of repeat 3-2, and again in repeat 5-1. This mosaic distribution can be considered as evidence for multiple crossing-over or conversion events between slightly different genes. The evolution of the downstream related region is far from clear. The similarity of repeats 4-1 and 5-1 and repeats 1-1, 2-1, and 3-1 strongly suggests that they arose from an ancient 150-bp element which existed prior to the amplification of the 250-bp repeat. Repeats 4-1 and 5-1 in Fig. 2b thus correspond to elements 5′ and 6′ in Fig. 3.

Frequent Elements

It can be seen from these results that certain sequence motifs are present many times in the rice IGS. The 150-bp element itself contains five copies of the GGG(G/C)(T/C)G motif, which is therefore repeated many times in the large repeat region. The (G/A)TTCGGTC motif which flanks the repeats is present 14 times within the repeat region, while the GAAAAC sequence can be found eight times in the repeats and, in contrast to the former sequence, is also present once upstream from the first repeat and once downstream from repeat 5-1. This may represent a trace of the increasingly degenerate repeats which can be observed downstream from the four repeated regions described above.

One of the rare sequence elements which is conserved between all species in the IGS is the site of transcription initiation. In higher plants it has been accurately determined in maize and wheat as well as in several dicots. It is often located downstream

```
RICE 1          CTGGACGCTAGGGGTGCGTGGGGCTGGCTA
RICE 2          CTGGACGCTAGGGGTGCGTTGGGCTGGCTA
RICE 3          CTGGACGCTAGGGGTGCGTTGGGCTGGCTA
MAIZE           CaGaACtCcccGaagtCcTGGGaggGGGTA
WHEAT           CT-aAgGgcAGGGGTCCcaatGGg-GGCTA
RYE             CTGaAgGgcAGGGGTCCaaaGGGg-GGCTA

RICE 1          TGGCCCACGACTATAGTAGGGGGGAAGGGA
RICE 2          TGGCCCTAGACTATAGTAGGGGGGAAGGGA
RICE 3          TGGCCCTAGCCTATAGTAGGGGTGAGCGGA
RICE 4          GGCACGTAGCCTATAccgGGccGGgGGGaG
MAIZE           TGCCCCTCaggTATAGTAGGGGGtAGGGaA
WHEAT           aaaCCCTCGggTATAGTAGGGaGGAGGGGt
RYE             aaaCCCTCGggTATAGTAGGGaGGAGGGGt

ARABIDOPSIS 1       TAAAGCTATA-TAGGGGTGGGT
ARABIDOPSIS 2       TTAAGCTATA-TAGGGGGGTGG
RADISH              TAGTGTTATA-TAGGGGGTAGG
MUSTARD             TAAGCATATATAAGGGGGGTAG
SPRING CABBAGE      TTTGTCTATATAAGGGGTAGGC
CUCUMBER            TAAGCATATATAAGGGGGGTAG
SQUASH 1            ATGTACTATA-TAGGGGGGTGA
SQUASH 2            ATGTACTATA-TAGGGGGGTGA
PUMPKIN 1           ATGTACTATA-TAGGGGGGTGA
PUMPKIN 2           ATGTACTATA-TAGGGGGGTGA
TOMATO              TAAGCATATATAAGGGGGGTAG
PEA                 CAAGCTTATA-TAGGGGGAGGC
CARROT 1            TCCATTCGTA-TGGAGGGGACC
CARROT 2            TCCATTCGTA-TGGAGGGGACC
CARROT 3            TCCATTCGTA-TGGAGGGGACC
MUNG BEAN 1         TACCCCCTTA-TAGGGGGGGAG
MUNG BEAN 2         CGAATATATA-TGGGGGGGGAG
MUNG BEAN 3         ATATCGAATA-TGCAGGGGGAG
BROAD BEAN          CAAACATATA-TAGGGGGAGGC
V. hirsuta          TTGATATATA-TAGGGGGGGGG
V. angustifolia     TAGCCATATATATGGGGGGACA
```

Fig. 4. Alignment of conserved sequences corresponding to putative transcription initiation sites. Sequences are aligned around the TATA(G)TA motif, the last A corresponding to the initiation site. Sequences of rice, maize, rye, and wheat present the extended homology upstream from the TATA sequence, whereas the other sequences show only the more limited region around the initiation site. *Dashes* indicate absent nucleotides. Bases which differ from the most frequently represented residue are shown in *lowercase*. Sequences are taken from rice (this paper), maize (McMullen et al. 1986), wheat (Barker et al. 1988), rye (Appels et al. 1986), *Arabidopdsis* (Gruendler et al. 1989), radish (Delcasso-Tremousaygue et al. 1988), mustard (Rathgeber and Capesius 1990), spring cabbage (Tremousaygue et al. 1992), cucumber (Zentgraf et al. 1990), squash (Kelly and Siegel (1989), EMBL accession number X13059), pumpkin (Zentgraf et al. 1990), tomato (Scmidt-Puchta et al. 1989), pea (Kato et al. 1990), carrot (Taira et al. 1988), mung bean (Gerstner et al. 1988), *Vicia hirsuta* (Yakura and Mishikawa (unpublished), EMBL accession number X62122), and *V. angustifolia* (Ueki et al. 1992).

from the repeats, at a variable distance from the 18S RNA coding region, while the repeating elements have been shown in certain species to play a role of transcriptional enhancer (Sollner-Webb and Tower 1986). Surprisingly, the homologous sequence in the rice IGS is found in each of the three repeats and is also present in a modified form in the related downstream, incomplete repetition. The alignment of the four potential transcription initiation sites in the repeats with other known or deduced initiation sites shows that they are highly conserved compared with these other sequences (Fig. 4). When the surrounding sequences are compared between rice, wheat, and maize, it is observed that the homology extends further upstream for about 25 bp and downstream for about 12 bp, suggesting that these ca. 60

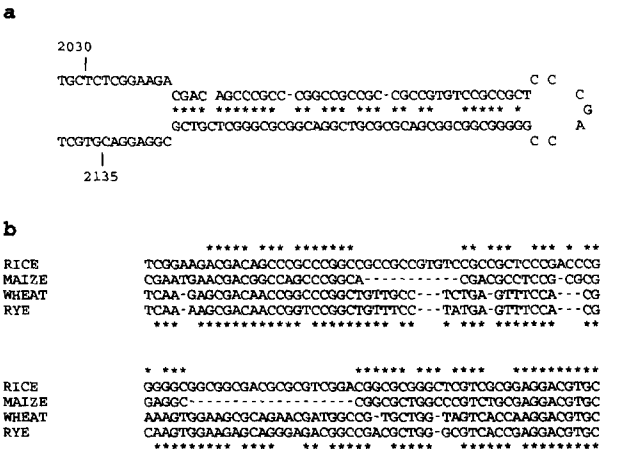


Fig. 5. Region of dyad symmetry upstream from the 18S rRNA coding region. **a** The region of the rice IGS immediately upstream from the 18S rRNA coding region is presented as a hairpin structure in which base pairing has been optimized. *Asterisks* show complementary bases. **b** Alignment of the rise region of dyad symmetry with those of maize, rye, and wheat. *Asterisks* above the alignment indicate identical bases between rice and maize sequences; those below, bases which are identical in rye and wheat. Coordinates are: rice (2,032–2,140), maize (3,046–3,118), wheat (4,001–4,094), and rye (4,151–4,245).

bp which constitute the central part of the 150 bp element have a common origin.

Potential Secondary Structures

Secondary structure analysis over the entire IGS region shows the presence of a certain number of regions of dyad symmetry. As in several other organisms (Barker et al. 1988), there is approximately 100-bp potential stem-loop structure in the transcribed region immediately upstream from the rice 18S RNA coding region (Fig. 5a) which is bordered by almost perfect direct repeats of 9nt (CGGA(A/G)GACG). A comparison of this sequence with similar structures in wheat, rye, and maize (Fig. 5b) shows that these sequences all derive from the same original sequence but have undergone a number of deletions since the different species diverged. The closer relationships between wheat and rye on the one hand and rice and maize on the other are clearly visible from this alignment. In wheat this hairpin structure is not in the immediate vicinity of the 18S rRNA but a few hundred base pairs upstream. In all cases the flanking sequences are imperfect duplications, reinforcing the hypothesis that this region may have been inserted as a transposable element (Barker et al. 1988).

Further regions of dyad symmetry are found within the repeats, one at the level of each of the sequences showing homology to the transcription initiation site and one flanking the GAAAAC se-

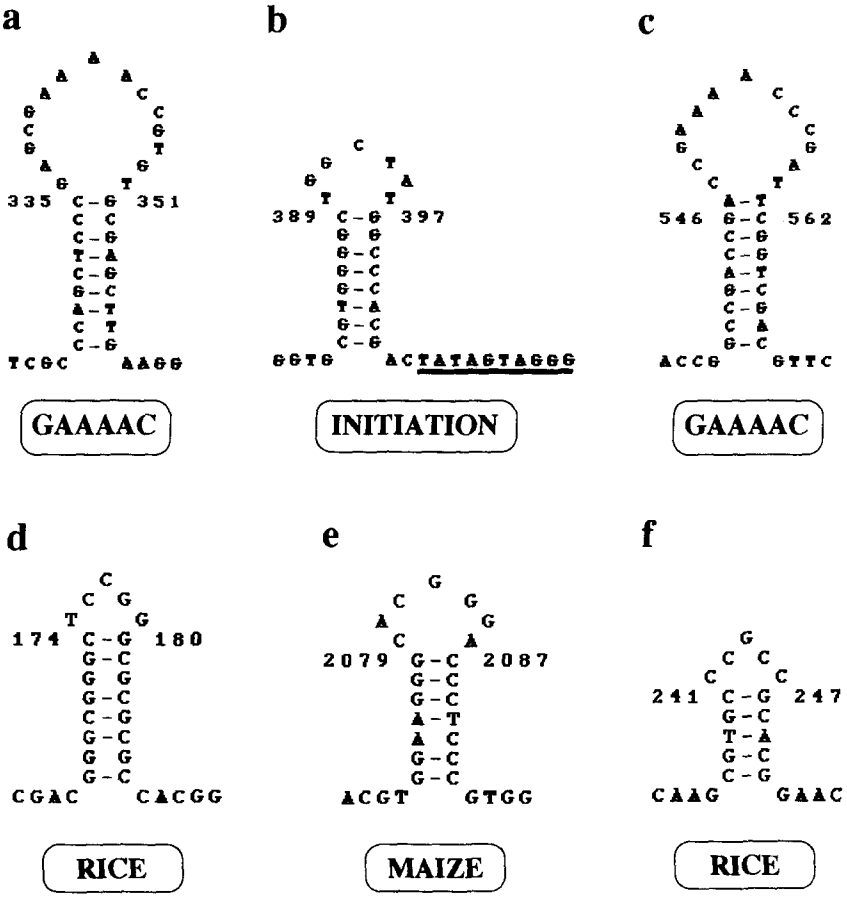


Fig. 6. **a–c** Potential secondary structures in the rice direct repeats. The sequences shown are from the first repeat, coordinates 305–569. The region around the transcription initiation site is underlined. **d–f** Sequence and structural similarities in the rice and maize spacers. The maize sequence corresponds to the putative transcription termination site downstream from the 25S rRNA coding region. Sequences are presented as stem-loop structures, as proposed by Barker et al. (1988). Coordinates of the sequences are indicated.

quences. The structures of these sequences in the first repeat, presented in a stem-loop form, are shown in Fig. 6a–c. In the former case, the structure is similar to those described in pea as well as in several other organisms (Piller et al. 1990). In the latter case, the GAAAAC sequences are found in the loop when these regions are represented as stem-loop structures.

Comparison with Other Cereal IGS Sequences

One of the striking features of ribosomal units is the lack of sequence conservation within the spacer sequences, even between closely related organisms, while the coding regions are almost perfectly conserved among widely diverging species. In order to study the overall structure of the rice spacer compared with the other cereal IGS sequences, we used a dot-matrix program which allows the detection of similar regions within a window of user-defined size and is particularly useful for the preliminary analysis of long sequences like the ribosomal spacers. It should, however, be borne in mind that this kind of comparison does not take into account gaps introduced into one or other of the sequences. Thus a long window will detect lengthy regions of homology while a shorter window will lessen the effect of gaps but increase the probability of fortuitous alignments. Figure 7 shows alignments of the rice IGS with those of maize, rye, and wheat.

It is immediately apparent that the regions of similar sequence are confined to certain regions within the spacers. All four spacers have a long region of very high homology of about 400 bp, with very few deletions, upstream from the 18S rRNA coding region. This region contains the potential hairpin sequence which is present in all four species and which was described above. It probably indicates that the sequence in this region of the transcript is important. This homology is greater between rice and maize than between rice and the two other sequences. However, it should be noted that this region does not extend to the transcription initiation site and therefore species-specific sequences are present within the transcription unit, as already demonstrated for rice (Cordesse et al. 1992). The apparent lack of homology directly upstream from the 18S rRNA coding region between the rice sequence and those of rye and wheat is due to the presence of the regions of dyad symmetry mentioned above, which have similar sequences but differ in length.

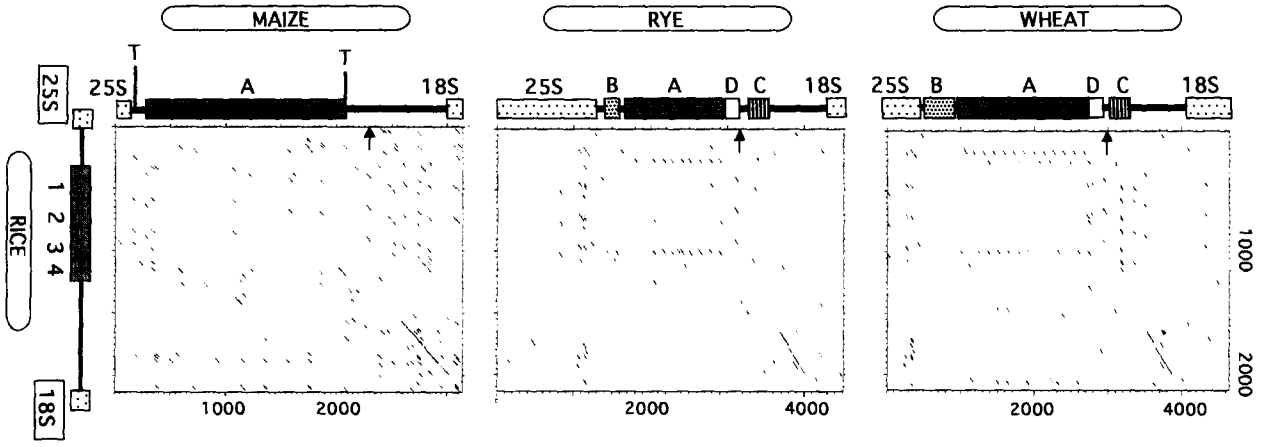


Fig. 7. Dot-matrix comparison of the rice IGS sequence with those of other cereals. Comparisons were carried out as described in Materials and Methods; 18S rRNA and 25S rRNA coding regions are indicated, as well as the different types of repeating units in each spacer. The *arrows* indicate transcription initiation sites. Sites on the maize spacer labeled T are the potential transcription termination sites (McMullen et al. 1986).

In the light of the presence of the potential transcription initiation sites within the rice repeats it is of interest to note that several sequence motifs which are outside the major repeat regions of the other spacers are present at least once in each of the rice repeats. In addition to the putative initiation sites, one motif in the 25S and another in the 18S rRNA coding region are present, while downstream from the repeats several elements can be seen, particularly in the wheat and maize sequences. One motif, which is present at the end of the first repeat, twice in repeats 2 and 3, and again in the truncated repeat 4 is homologous to the C repeats in the wheat IGS, which have been proposed to play a role in enhancing transcription (Barker et al. 1988). No homology can be seen between the rice repeats and the C repeats of rye, which are apparently not related to those of wheat.

In contrast to the limited alignments described above, one region of the dot matrices is practically devoid of homologous sequences. This is particularly noteworthy for the alignments of the rice IGS with those of wheat and rye where this region corresponds to the rice repeats, to the A and B repeats of wheat, and to the A repeats of rye. This demonstrates clearly that certain elements are highly species-specific. Nevertheless, it is interesting to note that individual repeats in wheat and rye are flanked by sequence motifs which can also be observed in the rice IGS. Examination of the sequence of the three spacers shows that there is a short region outside the rice repeats and directly upstream from the first of these, GAGCACCCGCACC, which is similar to the sequence GCGTACACGGACC in each of the wheat repeats and to the homologous sequence in the rye IGS. The second motif, GGACCCGTGAACGGG, separated from the first by a short, direct repeat on the wheat genome, is found on the rice IGS in the second part of the 253–264-bp repeats (G(G/C)A(A/C)CCGTGC(A/G)CGAG).

It is surprising to note the lack of extensive similarity between the rice and maize sequences despite the fact that maize is more closely related to rice than wheat or rye. In contrast to these last two species, there is almost no sequence similarity within or around the repeats. A certain number of elements outside this region appear to be found in both rice and maize sequences, but their distribution is much less structured than that seen with the wheat and rye spacers. Nonetheless, several motifs appear to be common to the two IGS sequences, two of which are of particular interest. The first is downstream from the 25S coding region (Fig. 6d) and corresponds to the proposed transcription termination site in maize (Fig. 6e, McMullen et al. 1986). The sequence in the rice IGS is also found three times within the repeats. The second motif is at the end of the last maize repeat, corresponding to a second putative termination site just upstream from the initiation site (McMullen et al. 1986). Two copies can be found in rice, upstream from the 253–264-bp repeats (Fig. 6f) and within the downstream related region. All these sequences can be folded into a stem-loop structure, as proposed for wheat and maize (Barker et al. 1988). A short element present in the transcribed part of the rice spacer is present within the maize repeats as well as several-fold in the transcribed part of the maize spacer.

Conclusion

Our results confirm and extend the short report on the same rice ribosomal IGS presented recently (Takaiwa et al. 1990); excluding a few sequence discrepancies, the general organization is the same

with three complete repeats of 264 bp, 253 bp, and 253 bp. These are flanked on their 3′ side by other related sequences. The presence of these repetitions confirms our observations on rice intergenic spacer polymorphism which showed that the length of the IGS in different cultivated Asian rice species differs by multiples of 260 bp (Cordesse et al. 1990, 1992). On the other hand the sequence determined by both Takaiwa and our group has little homology with previously reported partial data concerning an indica variety (Nandabalan and Padayatti 1989).

The 264–253-bp repeat element can be further divided into 150-, 60-, and 40-bp motifs. This origin of the 264–253-bp repeats by successive deletions from a 150-bp ancestral sequence is in contrast to that proposed by Barker et al. (1988) for wheat. Barker et al. suggest that in that organism the repeats have arisen by duplication and amplification of a very short sequence. We feel this is unlikely in rice as it can clearly be seen that the ends of the three subrepeats are similar, despite the difference in length, and it is not possible to identify a shorter motif which could have given rise to the present-day sequence as in wheat. Despite this observation on the overall structure of the repeats, two short regions in the subrepeats are made up of much shorter motifs and have thus probably arisen by slippage replication, as suggested by Barker et al. (1988). This is the case for the GACC motifs which are present in all three regions of each repeat and it is also possible to detect many degenerate copies of a GGGCTGG motif, one of which coincides with the beginning of the deletion in the 60-bp motifs compared with the 150-bp regions.

Thorough analysis of changes between homologous elements indicates that these are clustered, suggesting they arose in the present structure following several crossing-over or gene-conversion events between slightly different genes. Although the occurrence of crossing-overs is strongly suggested by the variation in copy numbers of repeats within the rDNA of different accession our data are the first suggestive evidence at the sequence level to support this hypothesis. Indeed, the analysis of the spacer fine structure has allowed us to reconstruct a plausible evolutionary scenario.

Ribosomal RNA intergenic spacer sequences are highly variable between even closely related organisms. However, within the spacers of the four cereal IGS sequences we examined there is nonetheless a large number of conserved regions, and these may be repeated and dispersed, giving a highly variable and patchwork organization. For example, four potential transcription initiation site sequences are present in the rice rDNA spacer, but in the absence of information on the transcription of the rice genes, it is impossible to decide whether all sites are actually used. The presence of multiple potential transcription initiation sites is not unusual and has already been described for other organisms (Gerstner et al. 1988; Gruendler et al. 1989; Kelly and Siegel 1989).

Very little is known about rDNA transcription termination in plants (Schiebel et al. 1989; Zentgraf et al. 1990). In cereals, potential sites have only been identified by sequence comparison. However, it is noteworthy that among the rare regions of similarity between the rice and maize spacers are short sequences similar to those proposed to play a role in transcription termination in maize (McMullen et al. 1986), although in rice they are at different positions and in different copy number. When these sequences are compared with wheat (Barker et al. 1988) there is no sequence similarity but a dyad symmetry is well conserved between the three species (Fig. 6d–f).

The most striking region of homology is in the transcribed sequence upstream from the 18S coding region. This region is conserved over a 400-bp stretch, followed by the region of dyad symmetry directly upstream from the 18S sequence, and may be involved in RNA processing (Barker et al. 1988). However, all four spacers contain species-specific sequences between the transcription initiation site and the homologous 400-bp region. These sequences correspond to the C repeats in wheat and rye, but the lack of similarity between the maize and rice IGS in this region is surprising, as the initiation sites are identical and the rest of the sequence up to the 18S gene is also highly conserved. However, when we analyzed various rice genomes we also found the greatest specificity in that region, with short sequences specific for the AA genome (Cordesse et al. 1992). Although no homology can be detected in the transcribed region between the rice IGS and the wheat and rye C repeats, a sequence related to at least a part of the wheat C repeats is present once or twice in each rice repeat. In wheat, it has been proposed that motifs in the C repeats may play a role in enhancement of transcription (Barker et al. 1988). The presence of similar sequences in the rice repeats indicates that they may play a comparable role, as the repetitive elements have been shown to act as transcriptional enhancers in other organisms (Sollner-Webb and Tower 1986).

The results of the dot-matrix comparisons clearly show that the longest repeated regions are the most highly variable between species. While Flavell (1989) deduces from a dot-matrix comparison of the wheat and rye IGS sequences that the A repeats in these two organisms are not similar, our data in Fig. 7 indicate that the wheat and rye A repeats derive from a sequence element located just upstream

from the first rice repeat. The right-hand part of this sequence is repeated in the rye A repeat whereas the left-hand part is present in the wheat A repeat. On the other hand, the sequence of the maize repeats seems to be totally unrelated to any rice sequence. Rice and maize repeats have a totally different origin from the other two cereals.

Several other elements are highly conserved in all four species. This is the case for the transcription initiation site(s) and the majority of the transcribed sequences upstream from the 18S rRNA coding regions. The GAAAAC motif described above is repeated twice in each rice repeat, once in wheat and rye repeats, and twice or three times in those of maize. Similar sequences, GAAAAC, GAAAAT, or both, are present in all IGS sequences we have examined. In radish the GAAAAT motif is protected against DNase digestion in footprinting experiments using homologous nuclear extracts (Echeverria et al. 1992), indicating that it might be a functionally important element. Finally, a number of short motifs whose sequence and/or structure have been conserved are present on all four spacers. Barker et al. (1988) suggested that such conservation could be indicative of a functional role for these motifs and their presence on the rice IGS sequence reinforces this hypothesis.

Our observation that sequences to the A repeats of both wheat and rye flank the rice repeats and the presence of multiple copies of the potential transcription initiation site in rice suggests that at least part of the observed divergence between cereal intergenic spacers may be due to the amplification of different regions of the ancestral IGS sequence. This phenomenon, in combination with the insertion of external elements, which have themselves been amplified in some cases (e.g., the C repeats in wheat), associated with slippage replication and the accumulation of point mutations would thus explain the variation in length and sequence which is currently observed in the cereal ribosomal spacers. Current progress in the identification of protein factors implicated in rRNA synthesis and their interaction with the IGS should allow a more precise definition of the sequences essential for ribosomal gene expression.

Acknowledgments. This work was supported by grants from The Rockefeller Foundation and the Centre National de la Recherche Scientifique. The generous gift of plasmid RR217 by Dr. Takaiwa is gratefully acknowledged.

References

Appels R, Moran LB, Gustafson JP (1986) The structure of DNA from the rye (Secale cereale) nor rl locus and its behaviour in wheat backgrounds. Can J Genet Cytol 28:673–685

Barker RF, Harberd NP, Jarvis MG, Flavell RB (1988) Structure and evolution of the intergenic region of a ribosomal DNA repeat unit of wheat. J Mol Biol 201:1–17

Chen EY, Seeburg PH (1985) Supercoil sequencing: a fast and simple method for sequencing plasmid DNA. DNA 4:165–170

Cordesse F, Second G, Delseny M (1990) Ribosomal gene spacer length variability in cultivated and wild rice species. Theor Appl Genet 79:81–88

Cordesse F, Grellet F, Reddy AS, Delseny M (1992) Genome specificity of rDNA spacer fragments from Oryza sativa L. Theor Appl Genet 83:864–870

Delcasso-Tremousaygue D, Grellet F, Panabieres F, Ananiev ED, Delseny M (1988) Structural and transcriptional characterization of the external spacer of a ribosomal RNA nuclear gene from a higher plant. Eur J Biochem 172:767–776

Dover G (1986) Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet 2:159–165

Dover G, Flavell RB (1984) Molecular coevolution: DNA divergence and the maintenance of function. Cell 38:622–633

Echeverria M, Tremousaygue D, Delseny M (1992) A nuclear protein fraction binding to dA/dT-rich sequences upstream from the radish rDNA promoter. Plant J 2:211–219

Flavell RB (1989) Variation in structure and expression of ribosomal DNA loci in wheat. Genome 31:963–968

Gerstner J, Schiebel K, von Waldburg G, Hemleben V (1988) Complex organization of the length heterogeous 5'external spacer of mung bean (Vigna radiata) ribosomal DNA. Genome 30:723–733

Grellet F, Delcasso-Tremousaygue D, Delseny M (1989) Isolation and characterization of an unusual repeated sequence from the ribosomal intergenic spacer of the crucifer Sisymbrium irio. Plant Mol Biol 12:695–706

Grellet F, Delcasso D, Panabieres F, Delseny M (1986) Organisation and evolution of a higher plant alphoid-like satellite DNA sequence. J Mol Biol 187:495–507

Gruendler P, Unfried I, Pointner R, Schweizer D (1989) Nucleotide sequence of the 25S-18S ribosomal gene spacer from Arabidopsis thaliana. Nucleic Acids Res 17:6395–6396

Hemleben V, Ganal M, Gerstner J, Schiebel K, Torres RA (1988) Organization and length heterogenicity of plant ribosomal RNA. In: Kahl G (ed) Genes in architecture of the eukaryotic gene, pp 371–383. VCH verlagsgellschaft, Weinkeim, Federal Republic of Germany

Higgins DG, Sharp PM (1989) Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5:151–153

Kato A, Nakajima T, Yamashita J, Yamashita J, Tanifuji S (1990) The structure of the large spacer region of the rDNA in Vicia and Pisum sativum. Plant Mol Biol 14:983–993

Kelly RJ, Siegel A (1989) The Cucurbita maxima ribosomal DNA intergenic spacer has a complex structure. Gene 80:239–248

Lassner M, Dvorak J (1986) Preferential homogenization between adjacent and alternate subrepeats in wheat rDNA. Nucleic Acids Res 14:5499–5012

Lassner M, Anderson O, Dvorak J (1987) Hypervariation associated with a 12 nucleotide direct repeat and inferences on intergenomic homogenization of ribosomal RNA gene spacers based on the sequence of a clone from the wheat nor-d3 locus. Genome 29:770–777

McMullen MD, Hunter B, Philips RC, Rubenstein I (1986) The structure of the maize ribosomal DNA spacer region. Nucleic Acids Res 14:4953–4958

Maniatis T, Fritsch E, Sambrook J (1990) Molecular cloning: a laboratory manual. Cold Spring Harbor University Press, Cold Spring Harbor, NY

Marck C (1988) "DNA Strider": a "C" program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Res 16:1829–1836

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564

Nandabalan K, Padayatty JD (1989) Initiation of transcription of rDNA in rice. Biochem Biophys Res Comm 160:1117–1123

Piller KJ, Baerson S, Polans NO, Kaufman LS (1990) Structural analysis of the short length ribosomal DNA variant from Pisum sativum L cv alaska. Nucleic Acids Res 18:3135–3145

Rathgeber J, Capesius I (1990) Nucleotide sequence of the intergenic spacer and the 18S ribosomal RNA gene from mustard (Sinapis alba). Nucleic Acids Res 18:1288–1288

Rogers JO, Bendich AJ (1987) Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. Plant Mol Biol 9:509–520

Sanger F, Nicklens S, Coulson AR (1977) DNA sequencing with chain termination inhibitors. Proc Natl Acad Sci USA 74: 7463–7467

Sano Y, Sano T (1990) Variation of the intergenic spacer region of ribosomal DNA in cultivated and wild rice species. Genome 33:209–218

Schiebel K, Von Waldburg G, Gerstner J, Hemleben V (1989) Termination of transcription of ribosomal RNA genes of mung bean occurs within a 175 bp repetitive element of the spacer region. Mol Gen Genet 218:302–307

Schmidt-Puchta W, Guenther I, Saenger HL (1989) Nucleotide sequence of the intergenic spacer (IGS) of the tomato ribosomal DNA. Plant Mol Biol 13:251–253

Sollner-Webb B, Tower J (1986) Transcription of cloned eukaryotic ribosomal RNA genes. Ann Rev Biochem 55:801–830

Staden R (1988) Methods to define and locate patterns of motifs in sequences. CABIOS 4:53–60

Taira T, Kato A, Tanifuji S (1988) Difference between two major size classes of carrot rDNA repeating units is due to reiteration of sequences of about 460 bp in the large spacer. Mol Gen Genet 213:170–174

Takaiwa F, Oono K, Sugiura M (1984) The complete nucleotide sequence of a rice 17S rRNA gene. Nucleic Acids Res 12: 5369–5379

Takaiwa F, Ikuchi S, Oono K (1990) The complete nucleotide sequence of the intergenic spacer between 25S and 17S rDNAs in rice. Plant Mol Biol 15:933–935

Toloczyki C, Feix G (1986) Occurrence of nine homologous repeat units in the external spacer of a nuclear maize rRNA gene unit. Nucleic Acids Res 14:4969–4986

Tremousaygue D, Laudie M, Grellet F, Delseny M (1992) The Brassica oleracea rDNA spacer revisited. Plant Mol Biol 18: 1013–1018

Ueki M, Uchizawa E, Yakura K (1992) The nucleotide sequence of the rDNA intergenic spacer region in a wild species of the genus vicia, V angustifolia. Plant Mol Biol 18:175–178

Van't Hof J, Lamm SS (1991) Single stranded replication intermediates of ribosomal DNA replicons of pea. EMBO J 10: 1949–1953

Yanisch-Perron C, Viera J, Messing J (1985) Improved m13 phage cloning vectors and host strains: nucleotide sequences of the m13mp18 and puc19 vectors. Gene 33:103–119

Zentgraf U, Ganal M, Hemleben V (1990) Length heterogeneity of the rRNA precursor in cucumber (Cucumis sativus). Plant Mol Biol 15:465–474

Received July 16, 1992/Revised September 18, 1992