The spliceosomal snRNAs of *Caenorhabditis elegans*

Jeffrey Thomas, Kristi Lea, Erin Zucker-Aprison and Thomas Blumenthal*
Department of Biology and Program in Molecular, Cellular and Developmental Biology, Indiana University, Bloomington, IN 47405, USA

Received February 1, 1990; Revised and Accepted April 2, 1990 EMBL accession nos X51371 – X51388 (incl.)

ABSTRACT

Nematodes are the only group of organisms in which both cis- and trans-splicing of nuclear mRNAs are known to occur. Most *Caenorhabditis elegans* introns are exceptionally short, often only 50 bases long. The consensus donor and acceptor splice site sequences found in other animals are used for both cis- and trans-splicing. In order to identify the machinery required for these splicing events, we have characterized the *C. elegans* snRNAs. They are similar in sequence and structure to those characterized in other organisms, and several sequence variations discovered in the nematode snRNAs provide support for previously proposed structure models. The *C. elegans* snRNAs are encoded by gene families. We report here the sequences of many of these genes. We find a highly conserved sequence, the proximal sequence element (PSE), about 65 bp upstream of all 21 snRNA genes thus far sequenced, including the SL RNA genes, which specify the snRNAs that provide the 5′ exons in trans-splicing. The sequence of the *C. elegans* PSE is distinct from PSE's from other organisms.

INTRODUCTION

The splicing of nuclear mRNAs is catalyzed by small, nuclear ribonucleoproteins (snRNPs) and associated proteins[1–3]. Four snRNPs have been shown to be involved in splicing; U1, U2, U5, and the complex of U4 and U6. The snRNAs U1, U2, U4, and U5, all possess a sequence of the motif 5′-purine-A-U_{4-6}-G-purine-3′, to which a group of proteins known collectively as 'Sm', binds[4,5]. These RNAs are also capped with 2,2,7-trimethylguanosine (TMG) at their 5′ ends.

The RNA component of U1 (the U1 snRNA) has been shown to interact with the 5′ splice-site by Watson-Crick base pairing[6]. The U2 snRNA interacts similarly with the branch site in *Saccharomyces cerevisiae*[7]. U5 snRNP appears to act at the 3′ splice acceptor[3]. U4/U6 has been shown to be involved in the assembly of the active splicing complex[3,8]. Several investigators have observed complexes containing both U4/U6 and U5[9,10]. However, at present, there is no indication of a specific role for U4/U6.

The U snRNAs are ubiquitously and abundantly expressed[11]. Most organisms studied have multiple copies of snRNA genes. However, the number of genes varies widely. The number of U1 genes, for instance, is only a few in plants and fruit flies, and about one-thousand in frogs. In mammals, there appears to be clustering of snRNA gene loci. The human U1 genes[12,13], human U2 genes[14], and mouse U1 genes[15] have been localized to single regions of the genome by *in situ* hybridization. However, *in situ* hybridization to *Drosophila* polytene chromosomes showed that the snRNA genes were scattered throughout the genome[16].

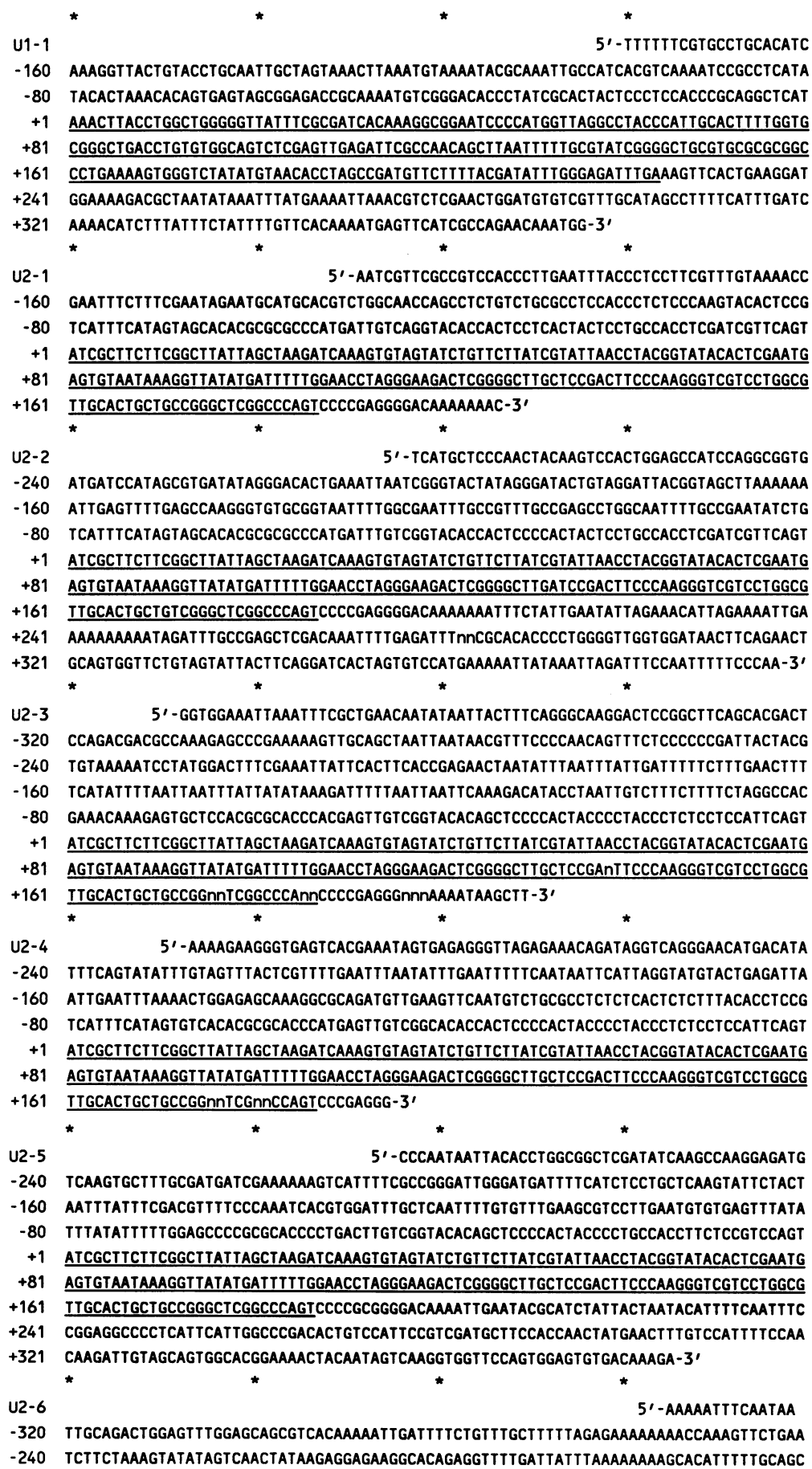
The U1, U2, U4, and U5 snRNA genes are transcribed by RNA polymerase II (pol II), based on the sensitivity of their activity to alpha-amanitin[5,11]. Analysis of vertebrate snRNA genes has revealed that their promoters differ from most mRNA promoters in several respects. First, these promoters lack TATA elements. Instead, an element known as the 'Proximal Sequence Element' (PSE), which is located about 40 to 70bp upstream of the transcription start site, specifies the site of transcription initiation and is essential for transcription. The sequence of the PSE is unrelated to the TATA box found about 30bp upstream of most mRNA genes. A second important sequence element, known as the 'Distal Sequence Element' (DSE), is located about 200bp upstream of the transcription initiation site. The DSE functions as an enhancer.

In contrast, the U6 snRNA gene is transcribed by RNA polymerase III[17]. However, its promoter exhibits some remarkable similarities to the other snRNA promoters and to mRNA promoters. Specifically, it contains a PSE which is required for U6 transcription and a TATA element typical in sequence and position to the standard mRNA TATA element.

It was recently discovered that in addition to standard (cis-) splicing, nematodes conduct a trans-splicing reaction in which a short, non-translated leader (called 'SL') is joined to the 5′ ends of a subset of their mRNAs[18,19]. 5′ and 3′ splice sites, typical in sequence of those found in standard *C. elegans* introns, are present at the joining sites of these trans-spliced RNAs. Hence, it is possible that snRNPs which are known to catalyze cis-splicing also participate in trans-splicing.

Since no nematode snRNPs had been characterized, we undertook the characterization of the *C. elegans* snRNPs. We have shown that all of the *C. elegans* spliceosomal snRNPs are Sm- bound and TMG-capped[20]. We report here the sequences and structures of the C. elegans snRNAs. The snRNAs are specified by small, dispersed gene families, some members of which we have cloned and localized on the nematode genome.

* To whom correspondence should be addressed


U1-1 5′-TTTTTTCGTGCCTGCACATC
-160 AAAGGTTACTGTACCTGCAATTGCTAGTAAACTTAAATGTAAAATACGCAAATTGCCATCACGTCAAAATCCGCCTCATA
-80 TACACTAAACACAGTGAGTAGCGGAGACCGCAAAATGTCGGGACACCCTATCGCACTACTCCCTCCACCCGCAGGCTCAT
+1 AAACTTACCTGGCTGGGGGTTATTTCGCGATCACAAAGGCGGAATCCCCATGGTTAGGCCTACCCATTGCACTTTTGGTG
+81 CGGGCTGACCTGTGTGGCAGTCTCGAGTTGAGATTCGCCAACAGCTTAATTTTTGCGTATCGGGGCTGCGTGCGCGCGGC
+161 CCTGAAAAGTGGGTCTATATGTAACACCTAGCCGATGTTCTTTTACGATATTTGGGAGATTTGAAAGTTCACTGAAGGAT
+241 GGAAAAGACGCTAATATAAATTTATGAAAATTAAACGTCTCGAACTGGATGTGTCGTTTGCATAGCCTTTTCATTTGATC
+321 AAAACATCTTTATTTCTATTTTGTTCACAAAATGAGTTCATCGCCAGAACAAATGG-3′
U2-1 5′-AATCGTTCGCCGTCCACCCTTGAATTTACCCTCCTTCGTTTGTAAAACC
-160 GAATTTCTTTCGAATAGAATGCATGCACGTCTGGCAACCAGCCTCTGTCTGCGCCTCCACCCTCTCCCAAGTACACTCCG
-80 TCATTTCATAGTAGCACACGCGCGCCCATGATTGTCAGGTACACCACTCCTCACTACTCCTGCCACCTCGATCGTTCAGT
+1 ATCGCTTCTTCGGCTTATTAGCTAAGATCAAAGTGTAGTATCTGTTCTTATCGTATTAACCTACGGTATACACTCGAATG
+81 AGTGTAATAAAGGTTATATGATTTTTGGAACCTAGGGAAGACTCGGGGCTTGCTCCGACTTCCCAAGGGTCGTCCTGGCG
+161 TTGCACTGCTGCCGGGCTCGGCCCAGTCCCCGAGGGGACAAAAAAAC-3′
U2-2 5′-TCATGCTCCCAACTACAAGTCCACTGGAGCCATCCAGGCGGTG
-240 ATGATCCATAGCGTGATATAGGGACACTGAAATTAATCGGGTACTATAGGGATACTGTAGGATTACGGTAGCTTAAAAAA
-160 ATTGAGTTTTGAGCCAAGGGTGTGCGGTAATTTTGGCGAATTTGCCGTTTGCCGAGCCTGGCAATTTTGCCGAATATCTG
-80 TCATTTCATAGTAGCACACGCGCGCCCATGATTTGTCGGTACACCACTCCCCACTACTCCTGCCACCTCGATCGTTCAGT
+1 ATCGCTTCTTCGGCTTATTAGCTAAGATCAAAGTGTAGTATCTGTTCTTATCGTATTAACCTACGGTATACACTCGAATG
+81 AGTGTAATAAAGGTTATATGATTTTTGGAACCTAGGGAAGACTCGGGGCTTGATCCGACTTCCCAAGGGTCGTCCTGGCG
+161 TTGCACTGCTGTCGGGCTCGGCCCAGTCCCCGAGGGGACAAAAAAATTTCTATTGAATATTAGAAACATTAGAAAATTGA
+241 AAAAAAAAATAGATTTGCCGAGCTCGACAAATTTTGAGATTTnnCGCACACCCCTGGGGTTGGTGGATAACTTCAGAACT
+321 GCAGTGGTTCTGTAGTATTACTTCAGGATCACTAGTGTCCATGAAAAATTATAAATTAGATTTCCAATTTTTCCCAA-3′
U2-3 5′-GGTGGAAATTAAATTTCGCTGAACAATATAATTACTTTCAGGGCAAGGACTCCGGCTTCAGCACGACT
-320 CCAGACGACGCCAAAGAGCCCGAAAAAGTTGCAGCTAATTAATAACGTTTCCCCAACAGTTTCTCCCCCCGATTACTACG
-240 TGTAAAAATCCTATGGACTTTCGAAATTATTCACTTCACCGAGAACTAATATTTAATTTATTGATTTTTCTTTGAACTTT
-160 TCATATTTTAATTAATTTATTATATAAAGATTTTTAATTAATTCAAAGACATACCTAATTGTCTTTCTTTTCTAGGCCAC
-80 GAAACAAAGAGTGCTCCACGCGCACCCACGAGTTGTCGGTACACAGCTCCCCACTACCCCTACCCTCTCCTCCATTCAGT
+1 ATCGCTTCTTCGGCTTATTAGCTAAGATCAAAGTGTAGTATCTGTTCTTATCGTATTAACCTACGGTATACACTCGAATG
+81 AGTGTAATAAAGGTTATATGATTTTTGGAACCTAGGGAAGACTCGGGGCTTGCTCCGAnTTCCCAAGGGTCGTCCTGGCG
+161 TTGCACTGCTGCCGGnnTCGGCCCAnnCCCCGAGGGnnnAAAATAAGCTT-3′
U2-4 5′-AAAAGAAGGGTGAGTCACGAAATAGTGAGAGGGTTAGAGAAACAGATAGGTCAGGGAACATGACATA
-240 TTTCAGTATATTTGTAGTTTACTCGTTTTGAATTTAATATTTGAATTTTTCAATAATTCATTAGGTATGTACTGAGATTA
-160 ATTGAATTTAAAACTGGAGAGCAAAGGCGCAGATGTTGAAGTTCAATGTCTGCGCCTCTCTCACTCTCTTTACACCTCCG
-80 TCATTTCATAGTGTCACACGCGCACCCATGAGTTGTCGGCACACCACTCCCCACTACCCCTACCCTCTCCTCCATTCAGT
+1 ATCGCTTCTTCGGCTTATTAGCTAAGATCAAAGTGTAGTATCTGTTCTTATCGTATTAACCTACGGTATACACTCGAATG
+81 AGTGTAATAAAGGTTATATGATTTTTGGAACCTAGGGAAGACTCGGGGCTTGCTCCGACTTCCCAAGGGTCGTCCTGGCG
+161 TTGCACTGCTGCCGGnnTCGnnCCAGTCCCGAGGG-3′
U2-5 5′-CCCAATAATTACACCTGGCGGCTCGATATCAAGCCAAGGAGATG
-240 TCAAGTGCTTTGCGATGATCGAAAAAAGTCATTTTCGCCGGGATTGGGATGATTTTCATCTCCTGCTCAAGTATTCTACT
-160 AATTTATTTCGACGTTTTCCCAAATCACGTGGATTTGCTCAATTTTGTGTTTGAAGCGTCCTTGAATGTGTGAGTTTATA
-80 TTTATATTTTTGGAGCCCCGCGCACCCCTGACTTGTCGGTACACAGCTCCCCACTACCCCTGCCACCTTCTCCGTCCAGT
+1 ATCGCTTCTTCGGCTTATTAGCTAAGATCAAAGTGTAGTATCTGTTCTTATCGTATTAACCTACGGTATACACTCGAATG
+81 AGTGTAATAAAGGTTATATGATTTTTGGAACCTAGGGAAGACTCGGGGCTTGCTCCGACTTCCCAAGGGTCGTCCTGGCG
+161 TTGCACTGCTGCCGGGCTCGGCCCAGTCCCCGCGGGGACAAAATTGAATACGCATCTATTACTAATACATTTTCAATTTC
+241 CGGAGGCCCCTCATTCATTGGCCCGACACTGTCCATTCCGTCGATGCTTCCACCAACTATGAACTTTGTCCATTTTCCAA
+321 CAAGATTGTAGCAGTGGCACGGAAAACTACAATAGTCAAGGTGGTTCCAGTGGAGTGTGACAAAGA-3′
U2-6 5′-AAAAATTTCAATAA
-320 TTGCAGACTGGAGTTTGGAGCAGCGTCACAAAAATTGATTTTCTGTTTGCTTTTTAGAGAAAAAAAACCAAAGTTCTGAA
-240 TCTTCTAAAGTATATAGTCAACTATAAGAGGAGAAGGCACAGAGGTTTTGATTATTTAAAAAAAAGCACATTTTTGCAGC

```
-160  TGAATTTCAGGCAACTTTGCTATCCCCGACTGACAAAAAGCGCGTAATTGGTACGCCAATGCACCATATCTTACGTGCAA
 -80  TGTTACTCTGCGACAGTACGCGCACCCCTGATTTGTCGGTACACCCGTCCGCACTACTCCTGCCACCTCCTCTATTCAGT
  +1  ATCGCTTCTTCGGCTTATTAGCTAAGATCAAAGTGTAGTATCTGTTCTTATCGTATTAACCTACGGTATACACTCGAATG
 +81  AGTGTAATAAAGGTTATATGATTTTTGGAACCTAGGGAAGACTCGGGGCTTGCTCCGACTTCCCAAGGGTCGTCCTGGCG
+161  TTGCACTGCTGCCGGGCTCGGCCCAGTGCCCGCGGGGACAAAATTAAGGGATAGTTGCATAATAAAACGTAATGGACAAA
+241  CGTTTTTTGAAAAATTATTGTTGAGTGTTCGCGCTTTGTTTTTTTTTATTGTAATAATTATTATAATAACAAGTTAAAAA
+321  AAAAGAGTTAAGTAATACCGACATCAGTTTTTCAGTCACGAGC-3'
      *                   *                   *                   *
U2-7                                                          5'-TGTATTTGGAGTT
-240  GGATGTAATTCAAGAGGTGCCAGAAGACTTCCCAGATGCTCAAGTTTGAATTTTTTGTTTTGTAATTGTTTTTTTTTAAA
-160  TTTCATTTTTTGTTTTATAAATTCATGTACGCCTGGCAACCAGTCTCTGTCTGCGCCTCCACCTTCTCCCCACACCTCTG
 -80  TCATTTCATAGTGACACACGCGCGCCCCTGATTTGTCGGTACACAGCTCCCGACTAGCCCTGCCGCCTCATCCATTCAGT
  +1  ATCGCTTCTTCGGCTTATTAGCTAAGATCAAAGTGTAGTATCTGTTCTTATCGTATTAACCTACGGTATACACTCGAATG
 +81  AGTGTAATAAAGGTTATATGATTTTTGGATCCTAGGAAAGACTCGGGGCTTGCTCCGACTTTCCGCGGGTCGTACTGGCG
+161  TTGCACTGCTGCCGGGCTCGGCCCAGTCCCCGAGGGGACAAATAAGCTT-3'
      *                   *                   *                   *
U2-8                5'-TGTAGAGGCATACCTTGCTAATTTTTATTTTGGAAACTATATATACATATACATATATATAT
-240  ATATATATATATATATGATGCGGCATGATGCGGCAAATTTGCCAAATTTGCCGTTTGCCGAATGACGTTTGTAAAACCGA
-160  ATTTCTTTCGAATCGAATGCATGCACGTCTTGCAACCATCCTCTGTCTGCGCCTCCACCCTCTTCCCACTTACACTCCGT
 -80  TATTTCATAGTAGCACACGCGCGCCCCCTGATTTGTCTGTACACCACTCCCCACTAGCCCTGGCACCTCGGTCGTTCAGT
  +1  ATCGCTTCTTCGGCTTATTAGCTAAGATCAAAGTGTAGTATCTGTTCTTATCGTATTAACCTACGGTATACACTCGAATG
 +81  AGTGTAATAAAGGTTATATGATTTTTGGAACCCAGGGAAGACTCGGGGCTTGCTCCGACTTCCCAAGGGTCGTCCTGGCG
+161  TTGCACTGCTGCCGGGCTCGGCCCAGTCCCCAAGGGGACAAACAAAAAAAAAATTGAAAGTGAATATTGAAAAAACAGATT
+241  TGCTGAGCTC-3'
      *                   *                   *                   *
U2-9  5'-CGTTCGAAATAGCTATCAATGTTACGGAAGACTGGCTCGGAGAGCAATTCCAACACTTCTGATGGCGAGAAACATTT
-160  GAGGTATACTAGGCCTGAAACTTCTGATATTAGTTAAGATGCGCCTTTAAGAAGTACAGTACTTCTGTCGCTTATTACAG
 -80  TCAACTTATAGTGCTCCACGCGCGCCCTGTATTTGTCGGTACACTGCCCCAAACCACACTAGCCACGAAAGTTTTTCAGT
  +1  ATCGCTTCTTCGGCTTATTAGCTAAGATCAAAGTGTAGTATCTGTTCTTATCGTATTAACCTACGGTATACACTCGAATG
 +81  AGTGTAATAAAGGTTATATGATTTTTGGAAACTAGGGAAGACTCGGGGCTTGCTCCGTCTTTCCAAGGGTCGTCCCGGCG
+161  TTGCACTGCTGCCGGGCTCGGCCCAGTCCCCGAGGGGGGTCAAAAAAGAAATACAGGCTGATTTAAAAAAGCAAAGAATT
+241  TCATTAGGAGCCAAAGTAATCGCTAGAAAACGGATGAAAATGTAGTTTTCAGATGGACCACCGCCTGTCTCACCACAACG
+321  TGCCCCATATTCTCGAG-3'
      *                   *                   *                   *
U4-1                                                          5'-TACTCGCTACATAAGTTCACTT
-160  AAATATGAAGTCTCATATATACAGTTACAGAAGAACGACGTCACTAGTAATAATATTGATTCTACTCCGCCCATGTTTAC
 -80  ACTATTTATGGCAAnGCGCGGAAACCGTCAAGTGCCCGTTCATTCAGTAAATGGTGCAAACTCGAGTGTACTCTATCTTT
  +1  AGCTTTGCGCTGGGGCGATAACGTGACCAATGAGGCTTTGCCGAGGTGCGTTTATTGCTGGTTGAAAACTTTTCCCAATT
 +81  GCCCGCGATGTCCCCTGAAACATGGGTGGCATACGCAATTTTTGAACnnCTCTAGGAGGCAGAAAATCTTATTATAAGAT
+161  CGGCTCTCATAATAAACTTGACATAAAATATGATATAAACATAGACGGAGGATATTATCATTTTCAACGTAGCTTATCAC
+241  ATCTTAACCGAATGTTCTCGAAATCTTATTTATTTATCGCAACCTTACTTCCCTTCCCTCTGTAATCCCATCGCAGTATT
+321  ATTTTCTCTCGTACAGGCCTTGTTGAGCAAATCATTGCCAATGTTGATACAGTTCCATGATACGGGATCAATATATATAC
+401  GAATGCGAATTTATCTGAAATCGAGCATGTTAACGGCTGATCTAGA-3'
      *                   *                   *                   *
U4-2       5'-TTTATAGAACACTCTATCCCTAGAACACTTTTACCCTTCTGTTGATATTGATTCAACCACGTCCCCGTTTAC
 -80  ATTAATTGTTGCAACGAGCGGAAACCACAAAATGTCGGGCCATCGAATGAAACCCCACAAATGTTTGAGTGACTATATTC
  +1  AGCTTTGCGCTGGGGCGATAACGTGACCAATGAGGCTTTGCCGAGGTGCGTTTATTGCTGGTTGAAAACTTTTCCCAATT
 +81  GCCCGCGATGTCCCCTGAAACATGGGTGGCATACGCAATTTTTGAACGCCTCTAGGAGGCAGAAAAACATCTTCAAAATT
+161  CATCTAAGTTCAnTTCGCATTTTCGATATTGCGAAGTTAATAAAACACAATATTTATACATACAAATATTCGTTGATAAG
+241  GTTTCTACCGATTGGCTTTGCATCATTGGCTGCTTGGGCAGCTAAGATCTTCTCCTTAGGAGTCAAT-3'
      *                   *                   *                   *
U4-3                                  5'-GAATTCGGAGGAACGAAACTTTACAGATCTTAAGAAAGTAGAAGAA
-240  AAATCAATTTTGAGATTTTTATTAGTTTTTATGCAGTTTTAATCGAAGGGAAGAGATTCGTTTCAAGAAAACAGAGAGAC
-160  TCCGATAAACCGCGAGCGATTGCAATTCGAAAAAGACGACCGCAACGCACACGCGCCGCATTCCGAAGAATACAGAATGC
 -80  AGAAAGCACCATTGTGCGCGAAAGCCGCGATTTGTCGGTTGCTGTGTTTTATAGGATCAGCTGAGTGAGCTGCTGAATTT
  +1  AGCTTTGCGCTGGGGCGATAACGTGACCAATGAGGCTTTGCCGAGGTGCGTTTATTGCTGGTTGAAAACTTTTCCCAATT
 +81  GCCCGCGATGTCCCCTGAAACATGGGTGGCATACGCAATTTTTGAACGCCTCTAGGAGGCAGAATATAAATACTGAACAG
```

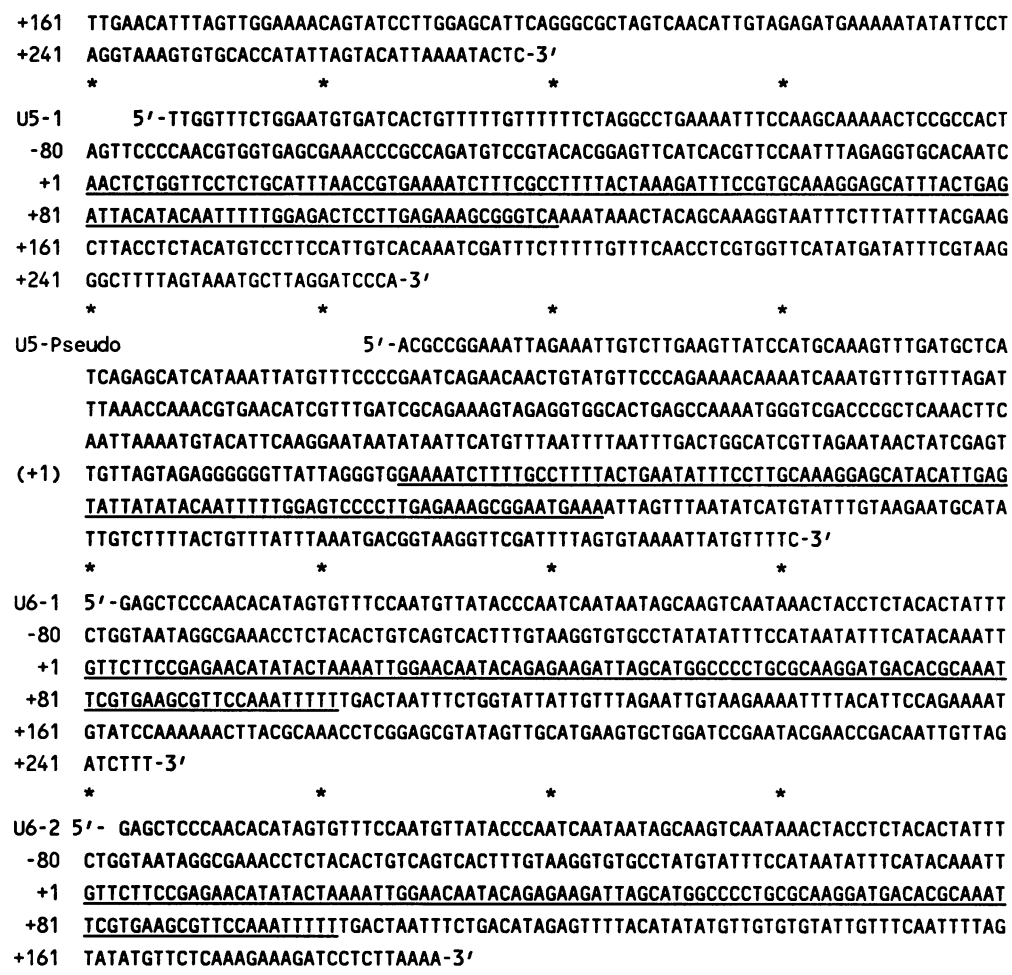


Figure 1. snRNA gene sequences. Appropiate fragments from the lambda clones shown in Fig. 2 were subcloned and sequenced by standard procedures. Genes are underlined. Numbering is from the first base of the snRNA. In the case of U5−PS, the region obviously related to U5 is underlined. U2-1a and U2-1b are presented as a single sequence, since there is only a single difference, at position −30.

Finally, we report the discovery of a putative PSE, present upstream of all 21 *C. elegans* snRNA genes thus far characterized.

MATERIALS AND METHODS

The *C. elegans* genomic library employed in this study was created by Karn et al[21]. Cloned phages were grown, maintained, and phage DNA prepared as described[20]. The hybridization probes used to clone the *C. elegans* snRNAs were 5'-^{32}P-labeled oligodeoxynucleotides. A list of these probes is provided in Thomas et al[20]. Oligonucleotides were labeled[22], purified electrophoretically or over Sephadex G-50 columns (Pharmacia) and hybridized to various blots as previously described[20]. Phage mapping and subcloning were performed using standard techniques. The pTZ vector system[23] was used exclusively.

C. elegans genomic DNA was prepared according to the method of Emmons et al[24]. In each restriction digest, ten micrograms of DNA was digested with 50 units of various restriction enzymes for 6−8 hrs. according to standard procedures. DNA was electrophoresed on 1% agarose gels and blotted onto Hybond-N using an alkaline blotting technique previously described[20]. Probes for these blots were DNA fragments generated by the polymerase chain reaction (PCR) and end-labeled with ^{32}P as described above for the labeling of oligonucleotides. PCR was performed using the GeneAmp DNA Amplification Reagent Kit from Perkin Elmer Cetus. PCR reactions were performed as described in the instructions provided by the manufacturer, except that nonidet-P 40 was added to a final concentration of 0.05%. Prehybridization was performed at 42°C for 2hr in 50% formamide, 6×SSC, 2.5×Denhardt's solution, 50mM sodium phosphate pH 6.5, 0.1% SDS, and 200μg/ml salmon sperm DNA. Hybridization was carried out in the same conditions overnight. Filters were washed twice for 15 min. in 2×SSC and 0.2% SDS at room temperature and then twice for 30 min. in 0.1×SSC and 0.1%SDS at 55°C.

Primer-extension sequencing of *C. elegans* snRNAs with Avian Myeloblastosis Virus reverse transcriptase (Seikagaku America) was performed as described by Geliebter[25].

RESULTS AND DISCUSSION

Cloning *C. elegans* snRNA Genes

C. elegans snRNA genes were cloned using radiolabeled oligodeoxynucleotides which were complementary to highly-conserved regions within previously characterized snRNAs. These oligonucleotides were capable of hybridizing to anti-Sm and anti-trimethylguanosine cap precipitable *C. elegans* RNAs

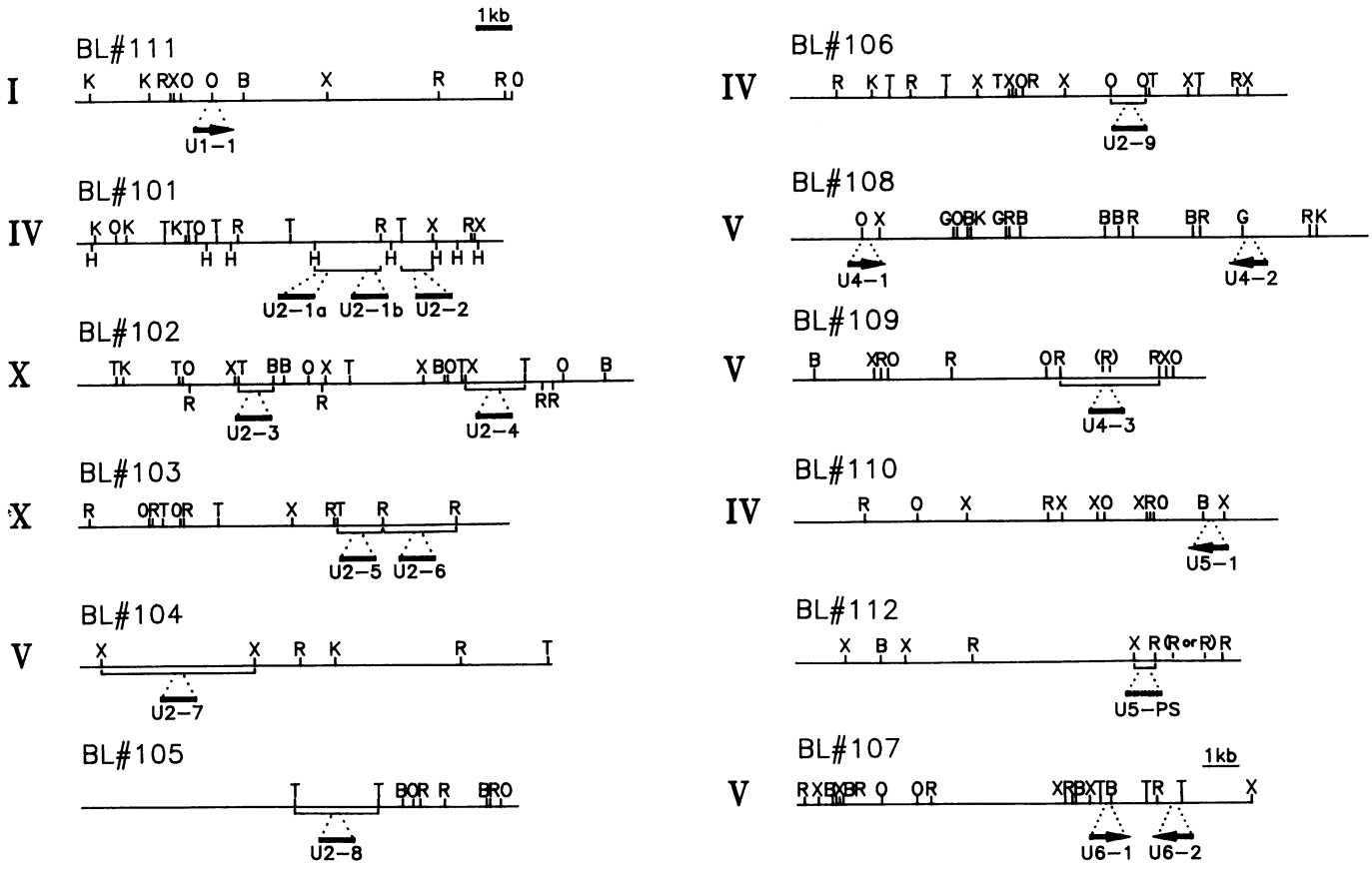


Figure 2. Restriction maps of *C. elegans* snRNA gene-containing clones. The location and orientation (where known) of the cloned snRNA genes are indicated. In cases where the exact position of the gene is unknown, a bracket denotes the restriction fragment harboring the gene. Restriction sites are indicated by the letters K, KpnI; R, EcoRI; X, XbaI; O, XhoI; B, BamHI; T, SstI; H, HindIII; G, BglII. Roman numerals indicate the chromosome onto which the phage has been located. The left end of the fragment shown abuts the left arm of the lambda host. BL # 112 contains the U5 pseudogene, as well as and EcoRI site which is present in one of the two locations shown. U1-2 and U1-3 are located on chromosome V, and a restriction map of these genes can be found in Russnak and Candido[26].

of the expected sizes (see Fig. 1 in ref. 20). After restriction mapping, subcloning and sequencing, the cloned genes were shown to encode snRNAs by comparing the DNA sequence to RNA sequence obtained from primer-extension of total *C. elegans* RNA (data not shown). Specifically, RNA sequence was obtained for nucleotides 15−48 of U1, 2-152 of U2, 2-41 of U4, 2-26 of U5, and 2-22 of U6 (see Fig. 4).

A single U1 gene, ten U2 genes, three U4 genes, a single U5 gene, and two U6 genes were cloned and sequenced including ~250bp upstream and ~100bp downstream of each gene (Fig. 1). In addition, a U5 pseudogene was sequenced. Restriction maps of snRNA gene-containing clones are shown in Fig. 2. Several clones contained pairs of snRNA genes (BL # 102, 103, 107, and 108), and one (BL # 101) contained three. The existence of two additional U1 genes, located at the borders of the *hsp-16* duplication[26], was also confirmed by primer-extension sequencing (not shown).

The locations of snRNA genes in the *C. elegans* genome

Several of the cloned genes have been localized onto the genomic map of *C. elegans* by molecular hybrization. These experiments were performed by Sulston and Coulson and by Lutterbach and Waterston (personal communications). The locations of these snRNA genes are denoted in Fig. 2. Since many of the phages isolated contained 2 or 3 genes, it is surprising that no further clustering of snRNA genes is evident. With the exception of the clustering noted above, the genomic organization of the cloned snRNA genes appears random.

Estimating snRNA gene number

Once subclones were obtained, genomic DNA blots were performed in order to estimate the size of the snRNA gene families. Due to the small size of the genes and the lack of useful restriction sites, we made probes using polymerase chain reaction in order to achieve specific hybridization. The results, shown in Fig. 3, clearly demonstrate that all of the *C. elegans* snRNAs are specified by small gene families. Some genomic DNA fragments produced more intense hybridization signals than others. There are several possible explanations for the differentially hybridizing fragments. First, there could be weak hybridization to related sequences, such as the pseudogene found on BL # 112 (Fig. 2), or to highly diverged snRNA genes. There could be strong hybridization to DNA fragments which contain multiple genes, of which there are several known examples (see Fig. 2). Finally, there could be comigration of gene-containing fragments. We estimate that there are about eleven U1 genes, twelve U2 genes, six U4 genes, nine U5 genes, and ten U6 genes.

Secondary structures of the *C. elegans* snRNAs

The existence of catalytic RNAs, and in particular of self-splicing RNAs, suggests that the RNA moieties of snRNPs are directly involved in catalysis of nuclear mRNA splicing. Hence, the

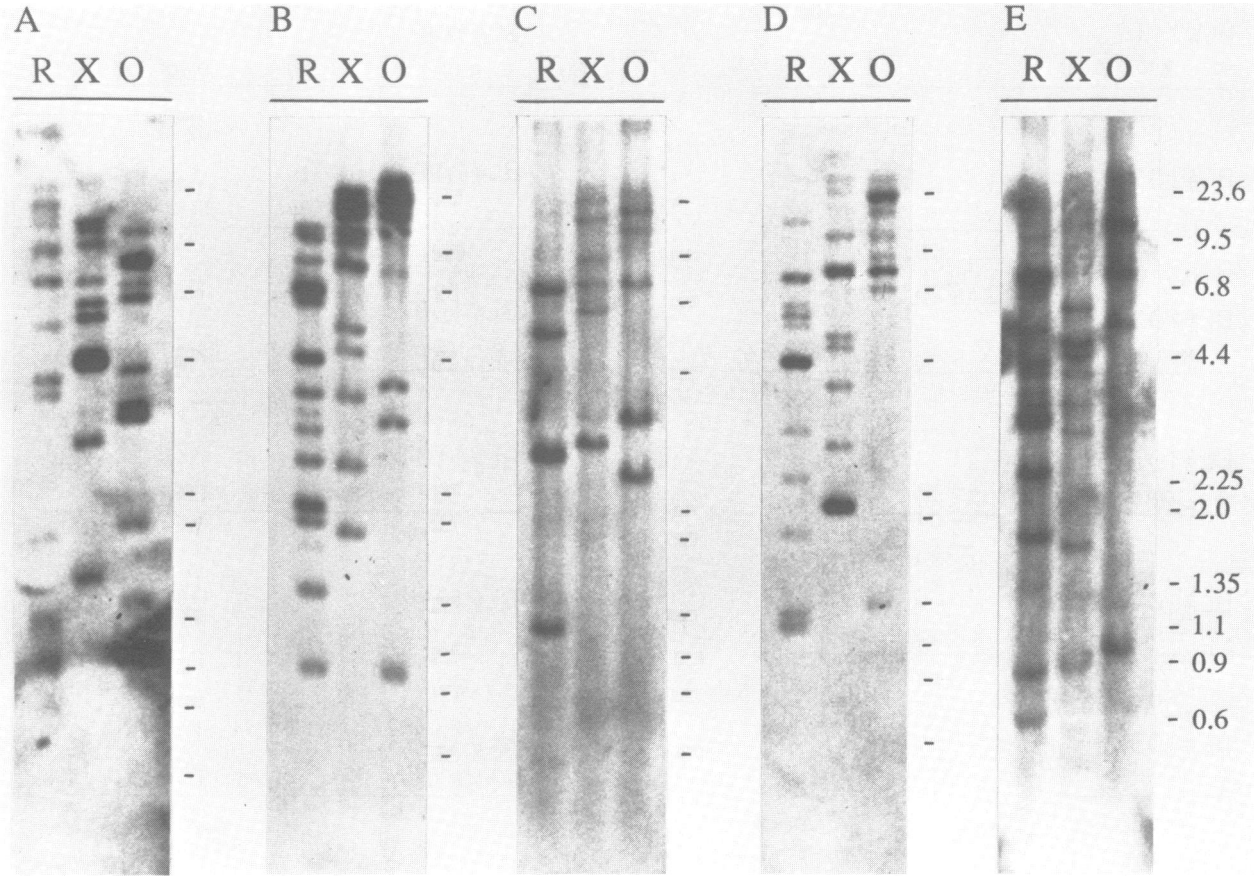


Figure 3. Genomic DNA blot analysis of snRNA genes. *C. elegans* genomic DNA was restricted with the enzymes EcoRI (R), XbaI (X) and XhoI (O), electrophoresed on a 1% agarose gel, and blotted onto Hybond-N. The blots were probed with end-labeled DNA fragments generated by polymerase chain-reaction (PCR). In A), the blot was probed with a fragment corresponding to −99 to +161 of U1-1 (relative to the start site of transcription). In B) the blot was probed with an end-labeled PCR-generated fragment which corresponds to −5 to +184 of U2. The blot in C) was probed with a fragment which corresponds to −20 to +140 of U4-1. In D) the blot was probed with a fragment corresponding to −92 to +121 of U5-1. The blot in E) was probed with a fragment corresponding to −49 to +98 of U6-1.

structure of the snRNAs is of great interest. The most informative and accurate approach toward inferring higher order RNA structure has proven to be phylogenetic comparison of homologous RNA sequences[27-29]. Since the sequences of snRNAs from many different organisms are known, several authors have performed such analyses[9,30,31] and phylogenetically-supported structure models exist for all five splicesomal RNAs.

We analyzed the structures of the *C. elegans* snRNAs for several reasons. First, conservation of important structural elements constitutes evidence that the true snRNA gene homologs have been cloned. Second, since no snRNAs from members of the nematode phylum have been characterized, the comparison of *C. elegans* snRNA could help to confirm or disprove proposed structure elements. Finally, such a phylogenetic comparison could help to reveal novel features of the *C. elegans* snRNAs—features which may function in accomodating trans-splicing.

Such a phylogenetic analysis was performed comparing the newly-acquired *C. elegans* snRNA sequences with the sequences of snRNAs from other organisms [both by us and by Guthrie and Patterson[9], to whom we provided the sequences of *C. elegans* U2, U4, and U6]. Our analysis was consistent with the results of Guthrie and Patterson. The minimal secondary structures of *C. elegans* spliceosomal snRNAs are shown in Fig. 4.

The first loop region of human U1 has been shown to contain the major recognition determinants for the U1-specific protein U1-70K[32,33]. This sequence is identical between *C. elegans*, *Drosophila*, *Xenopus*, and human U1. The hairpin region, known to interact with the U1-A protein[34], is also highly conserved between *C. elegans* and other organisms. Several compensatory nucleotide changes in the *C. elegans* sequence provide additional support for the helical elements in U1.

The *C. elegans* U2 sequence is particularly helpful in confirming proposed structural elements. In the region of U2 from position 7 to 56 (Fig. 4), *C. elegans* has more nucleotide differences from human U2 than do *Schizosaccharomyces pombe* or *S. cerevisiae*. A compensatory change (positions 15−20/G-C to U-A, also present in *Trypanosoma brucei* U2[35]) is the only phylogenetic evidence for the existence of the first stem. The branch site interaction region of U2 (nucleotides 35−40, see Parker et al[7]) is followed by a pseudoknot[30] comprised of nucleotides 49 through 54 and nucleotides 63 through 68 (Fig. 4). Three compensatory nucleotide changes prove the existence of this helical element, two of which are unique to *C. elegans* U2 (52−65/C-G, 53−64/G-C). Overall, the sequence and putative secondary structure of *C. elegans* U2 are quite similar to those characterized in other multicellular eukaryotes, and it possesses no unusual sequence or structural elements. In contrast, the U2 present in *T. brucei* (also known to conduct trans-splicing) is very unusual in the branch-site interaction region[35].

The intraspecies microheterogeneity of the *C. elegans* U2

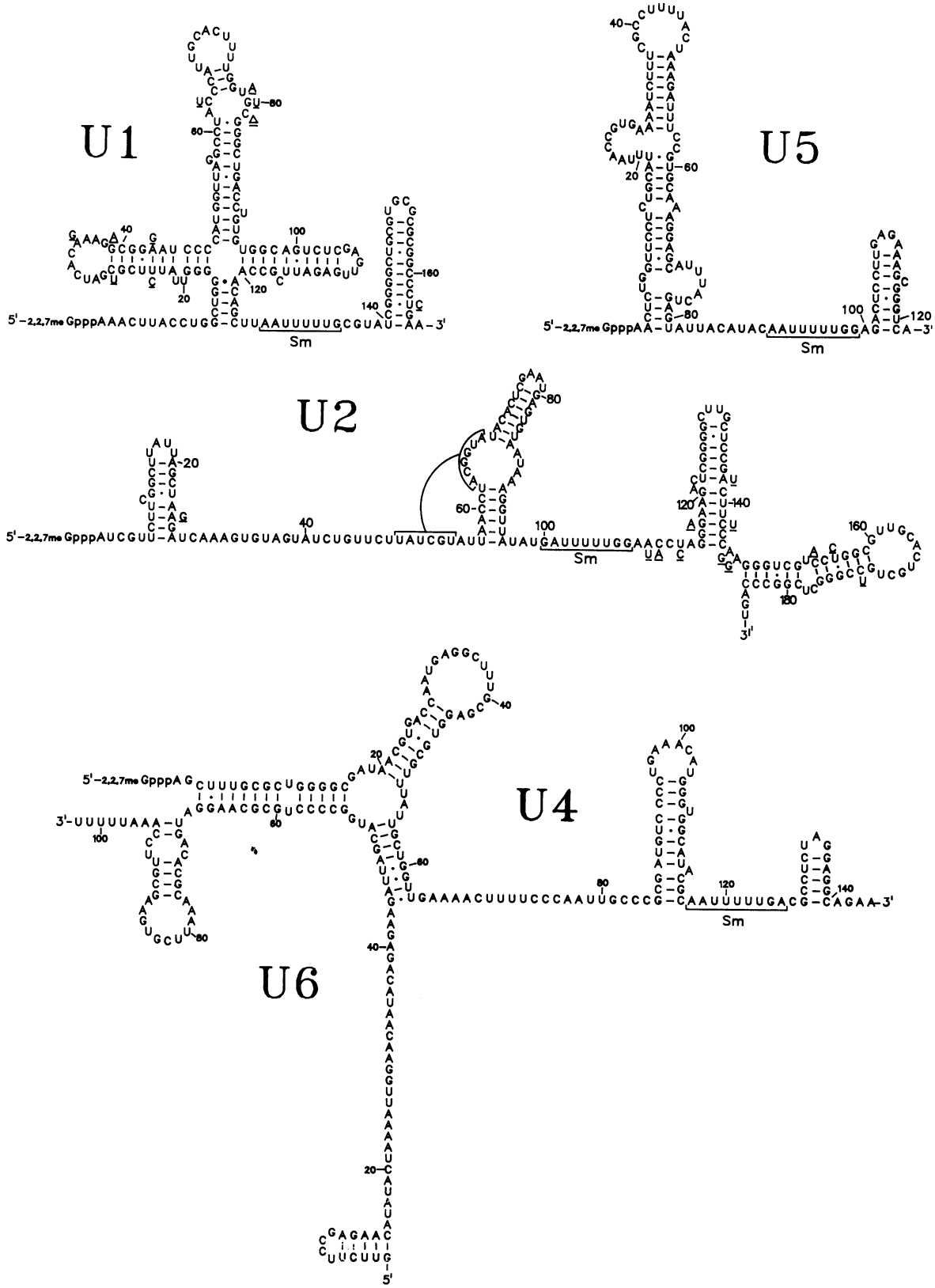


Figure 4. The *C. elegans* spliceosomal snRNAs. All snRNAs are shown in their most probable secondary structure, based on phylogenetic sequence comparison. Among the members of the U1 and U2 genes cloned to date, there is sequence variation. Alternate bases (underlined) are present next to their respective positions. Alternate bases shown in U1 are from U1-2 and U1-3 (which are identical in sequence). Except for a deletion at position 81, all of the changes are substitutions. The main U2 sequence represents the sequence o U2-1a, -1b, -3, -4, -5, and -6. U2-2 harbors a U at position 172. U2-7 has a U at position 110, an A at 117, a U at 142, a G at 145, a G at 146, and an A at 154. U2-8 contains a C at 113. U2-7 has a G at position 25, and A at 111, a U at 138, a U at 142, and a C at 156. The U4/U6 structure is from Zucker-Aprison et al[38].

```
                5'                              3'
              A * *     * *         * * * *  *
Consensus:    CGCGNAAACCNCNAAA--TGTCGGTACAC

U1-1     -61  AGCGGAGACCGCAAAA--TGTCGGGACAC
U1-2,3   -61  CGCGGAACCCGCCAGA--TGTCATGCTAT

U2-1a,b  -65  CACGCGCGCCCATGAT--TGTCGGTACAC
                                       ^A
U2-2     -65  CACGCGCGCCCATGATT-TGTCGGTACAC
U2-3     -65  CACGCGCACCCACGAGT-TGTCGGTACAC
U2-4     -65  CACGCGCACCCATGAGT-TGTCGGCACAC
U2-5     -65  CCCGCGCACCCCTGACT-TGTCGGTACAC
U2-6     -65  TACGCGCACCCCTGATT-TGTCGGTACAC
U2-7     -65  CACGCGCGCCCCTGATT-TGTCGGTACAC
U2-8     -65  CACGCGCGCCCCTGATT-TGTCTGTACAC
                      ^C
U2-9     -65  CACGCGCGCCCTGTATT-TGTCGGTACAC

U4-1     -63  CGCGGAAACCGTCAAG--TGCCCGTTCAT
U4-2     -63  AGCGGAAACCACAAAA--TGTCGGGCCAT
U4-3     -63  CGCGAAAGCCGCGATT--TGTCGGTTGCT

U5-1     -64  AGCGAAACCCGCCAGA--TGTCCGTACAC

SL-1     -66  ATCGTAGACCTCGCTCGCTGTCTGCTGTA
SL-1*    -69  GACGCGGGCCTCGATTGGTGTCGGCCGCT

SL-2     -65  CTCGCTTGCCTCGAAATGTGTCGGTACAC

U6-1,2   -72  GGCG-AAACCTCTACAC-TGTCAGT-CAC
```

Figure 5. The putative *C. elegans* proximal sequence element. The numbers to the right of the gene names indicate the position of the 5′-most nucleotide of the sequence shown, relative to the transcription initiation site. The consensus of all these sequences is indicated at the top, with especially highly conserved positions asterisked. SL-1* is from the SL RNA 1 gene from *Caenorhabditis briggsae* (P. Candido, personal communication). The SL-2 sequence is from Huang and Hirsh[43].

RNAs (shown in Fig. 4) parallels the interphylum heterogeneity of U2. Changes are clustered 3′ of the Sm binding site, in stems III and IV. Nearly all of the changes within the *C. elegans* U2 gene family are conservative with respect to secondary structure. The only non-conservative changes are the A to U change at position 138, and the C to A change at position 154. The former change is postulated to move a bulge one position 3′, and the latter change would increase the size of a bulge by one nucleotide in both strands of the helix in the model shown. Since the position and size of these bulges are not conserved (Fig. 4 and ref. 9), these differences also parallel the interphylum heterogeneity observed in U2.

The U5 snRNA is the least conserved of all of the spliceosomal snRNAs with respect to nucleotide sequence[9]. However, U5 seems to be highly-structured. By far the most conserved region is the uppermost part of stem/loop I (positions 38−48); the sequence in this region is identical in *C. elegans*, *Drosophila*, and human. Mutational analysis of human U5 suggests that this region is likely to mediate the association of U5 snRNA with the intron-binding protein[36,37]. *C. elegans* U5 is easily aligned with U5 RNAs of other organisms, and its pattern of conservation parallels the general pattern. However, since relatively few homologous nucleotides can be identified, it is difficult to assess whether *C. elegans* U5 has any unusual sequence elements.

The analysis of the structures of U4 and U6 snRNAs is unusual in that it must take into account the two intermolecular stems which are proposed to mediate the biological association of these snRNAs[31,38]. Phylogenetic evidence for their existence is especially important in light of experiments which show that the T_m of dissociation of these RNAs is the same whether U4/U6 snRNP is left intact or if it is treated with SDS and proteinase K[39]. The intermolecular stem I is so highly conserved that there are no compensatory changes which support its existence. Stem II is less conserved and its existence is proved by several compensatory changes in the yeast, trypanosome, and plant sequences. In general, the patterns of *C. elegans* U4 and U6 structure and sequence homology are not strikingly different from those in other organisms.

C. elegans snRNA Gene Promoters

In all snRNA genes characterized to date, the promoter elements are contained within the 5′-flanking sequence[5]. *C. elegans* snRNA genes do not appear to be an exception in this regard. Sequence similarities end shortly after the transcription unit ends, at which point the nucleotides are predominantly comprised of adenylic and thymidylic acids. Several regulatory elements have been defined in the snRNA genes of other organisms[40−44]. Except for two elements, the proximal sequence element (PSE) and the distal sequence element (DSE), these sequences are not conserved between different gene families or between different species examined[10]. A manual search of 5′ flanking regions from *C. elegans* U1, U2, U4, U5, and U6 snRNA genes not only failed to reveal any sequences similar to the gene- and species-specific elements discovered previously, but also failed to reveal any sequences resembling the PSE or DSE characterized in other organisms.

The *C. elegans* gene family in which the greatest number of 5′-flanking regions has been sequenced is the U2 gene family. We found a striking pattern of nucleotide conservation within these sequences. The region from −1 to −64 (relative to the transcription start site) is very similar in all 9 sequences; the nucleotides at 53 out of 64 positions are identical in seven out of the nine sequences (Fig. 1). In fact, so few positions are not conserved that discrete sequence elements within this region can not be defined. Upstream of that region, sequences become dissimilar and rich in adenine and thymine bases.

A comparison of sequences 5′ of the *C. elegans* U4 genes characterized is much less striking than that of the U2 genes. Relatively few nucleotide positions are conserved between the three putative U4 promoters sequenced. In only one region, centered around −50 (see below), is there a stretch of conserved positions. Since these regions are so dissimilar, more U4 promoter sequences must be obtained before such a sequence comparison will be informative.

The two cloned U6 genes differ at a single position of the 159bp of available upstream sequence, suggesting that they represent an evolutionarily recent duplication or conversion event. Thus a comparison of these sequences is of little value.

A comparison of all *C. elegans* snRNA gene upstream regions revealed a common sequence motif which is centered at about −50 relative to the transcription initiation site (Fig. 5). In the absence of a sequence resembling the PSE found in the snRNA promoters of other organisms, we presume this sequence represents the the *C. elegans* version of the PSE. The position of the proposed *C. elegans* PSE is similar to the position of the PSE's defined in other organisms, which are located between −40 and −70[5]. It is present not only in the U1, U2, U4, and U5 genes of *C. elegans* (presumably transcribed by pol II), but also in the U6 genes (presumably transcribed by pol III. The presence of the PSE in both of these gene classes is a further

similarity to PSE's in other organisms. Additionally, sequences conforming to the *C. elegans* PSE are present in a similar position in the 5'-flanking regions of the SL RNA genes in *C. elegans* [known to be transcribed by RNA polymerase II (Bektesh, Golomb, and Hirsh, personal communication)]. The element is 26 to 29 nucleotides long, depending on the gene. In comparison to other promoter elements characterized, the *C. elegans* PSE is unusually long, suggesting that two transcription factors may bind to it. Since there is no obvious similarity between the 5' and 3' halves of this element, these would likely be two different factors.

We note that some promoter regions contain much better matches to the consensus than do others. The U1, U5, and SL2 genes contain the best matches, while SL1 RNA genes are the most different. Perhaps this is because a reduced affinity for the transcription factor(s) is necessary to prevent overproduction of SL1 RNA, since the SL1 RNA genes are highly reiterated[18].

The only snRNA genes from other nematodes to be reported to date are the SL1 RNA genes of *Ascaris suum* and *Brugia malayii*. Interestingly, these genes do not contain a sequence related to the proposed *C. elegans* PSE. An *in vitro* system derived from *Ascaris* that accurately transcribes the *Ascaris* SL1 gene[46], fails to transcribe the *C. elegans* snRNA genes (Maroney, Hannon and Nilsen, personal communication). Thus it appears that the snRNA gene controlling elements have diverged surprisingly far within the nematode phylum.

Transcription Termination Signals in *C. elegans* snRNA Genes

The formation of mature 3' ends of vertebrate U1, U2, and U4 have been shown to be post-transcriptional events. Transcription termination occurs up to 15 nucleotides downstream of the mature 3' ends of the snRNAs[41]; the extra nucleotides are probably trimmed by an exonuclease[42,47]. Transcription termination requires three sequence elements; 1) sequences within the promoter[47,48] 2) snRNA-internal sequences[41,48,49] and 3) a sequence element 3' of the transcription unit (the 3' box)[41,47]. The best defined of these elements is the 3' box. Yuo et al[47] found sequences similar to the human 3' box in snRNA genes from several organisms, all within 35 nucleotides of the base corresponding to the mature 3'-end of the snRNA. They derived a consensus sequence for the 3' box: $GTTTN_{0-3}AAAG/ANNAGA$ (G/A signifies a single nucleotide position and indicates that either G or A is present there). No sequence resembling the 3' box can be found 3' of the *C. elegans* U1, U2, U4 or U5 genes. There is a conserved sequence present just 3' of the U2 genes which is capable of forming a stem-loop structure. Otherwise, there are no blocks of conserved sequences in these 3' regions, except for several blocks of A- and T-rich sequences.

U6 genes are transcribed by pol III. The termination of transcription of U6 genes is mediated by a run of thymidine residues at the 3' ends of the genes[5]. A run of thymidines is present in the *C. elegans* U6 genes starting at position +98.

ACKNOWLEDGEMENTS

We are grateful to Christine Guthrie and JoAnn Wise for gifts of oligonucleotides and to Susan Bektesh, Meriam Golomb, David Hirsh and Tim Nilsen for sharing their unpublished results. We are also grateful to Wolf Gallwitz for his help with DNA sequencing. This work was supported by grant GM30870 to T.B. from the National Institute of General Medical Sciences. J.D.T. was supported by a Research Service Award GM7227 from N.I.G.M.S.

REFERENCES

1. Padgett,R.A., Grabowski,P.J., Konarska,M.M., Seiler,S., and Sharp,P.A. (1986) *Annu. Rev. Biochem.* **55**, 1119–1150.
2. Maniatis,T. and Reed,R. (1987). *Nature* **325**, 673–678.
3. Steitz,J.A., Black,D.L., Gerke,V., Parker K.A., Kramer, A., Frendewey,D. and Keller,W. (1988) in Birnsteil,M.L. (ed.), Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles. Springer-Verlag, New York, 115–154.
4. Riedel,N., Wolin,S., and Guthrie,C. (1987) *Science* **235**, 328–331.
5. Parry,H.D., Scherly,D. and Mattaj,I.W. (1989) *Trends Biochem. Sci.*, **14**, 15–19.
6. Zhuang,Y., and Weiner,A. (1986) *Cell* **46**, 827–835.
7. Parker,R., Siliciano,P.G., and Guthrie,C. (1987) *Cell* **49**, 229–239.
8. Grabowski,P.J. and Sharp,P.A. (1986) *Science* **233**, 1294–1299.
9. Guthrie,C. and Patterson,B. (1988). *Annu. Rev. Gen.* **22**, 387–419.
10. Konarska,M.M. and Sharp,P.A. (1988) *Proc. Natl. Acad Sci. USA* **85**, 5459–5462
11. Dahlberg,J.E. and Lund,E. (1988) in Birnsteil,M.L. (ed.), Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles. Springer-Verlag, New York,
12. Naylor,S.L., Zabel,B.U., Manser,T., Gesteland,R. and Sakaguchi,A.Y. (1984) *Som. Cell Mol. Gen.* **10**, 307–313.
13. Lindgren,V., Bernstein,L.B., Weiner,A.M., and Francke,U. (1985) *Mol. Cell. Biol.*, **5**, 2172–2180.
14. Lindgren,V., Ares,M., Weiner,A.M. and Francke,U. (1985) *Nature* **314**, 115–116.
15. Blatt,C., Saxe,D., Marzuluff,W.F., Lobo,S., Nesbitt,M. and Simon,M.I. (1988) *Som. Cell Mol. Gen.* **14**, 133–142.
16. Alonzo,A., Jorcano,J.L., Beck,E., Hovemann,B. and Schmidt,T. (1984) *J. Mol. Biol.* **180**, 825–836.
17. Sollner-Webb,B. (1988). *Cell* **52**, 153–154.
18. Krause,M., and Hirsh,D. (1987) *Cell* **49**, 753–761.
19. Blumenthal,T., and Thomas,J.D. (1988) *Trends Genetics* **4**, 305–308.
20. Thomas,J.D., Conrad,R.C. and Blumenthal,T. (1988) *Cell* **54**, 533–539.
21. Karn,J., Brenner,S., Barnett,L. and Cesareni,G. (1980) *Proc. Natl. Acad. Sci.*, U.S.A. **77**, 5172–5176.
22. Maniatis,T., Fitsch,E.F., and Sambrook,J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor University Press, Cold Spring Harbor.
23. Mead,D.A., Szczesna-Skorupa,E. and Kemper,B. (1986) *Protein Eng.* **1**, 67–74.
24. Emmons,S.W., Klass,M.R. and Hirsh,D. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1333–1337.
25. Geleibter,J. (1986) *BRL Focus* **9**, 5–8.
26. Russnak,R.H. and Candido,P.M. (1985) *Mol. Cell. Biol.* 5, 1268–1278
27. Fox,G.E. and Woese,C.R. (1975) *Nature* **256**, 505–507.
28. Noller,H.F. and Woese,C.R. (1981). *Science* **212**, 403–411.
29. James,B.D., Olsen,G.J., Liu,J. and Pace,N.R. (1988) *Cell* **52**, 19–26.
30. Ares,M. and Igel,A.H., (1988) in Cech,T.(ed.) Molecular Biology of RNA, UCLA Symposia on Molecular and Cellular Biology. Alan R. Liss, New York **94**, 13–23.
31. Brow,D.A., and Guthrie,C.G. (1988) *Nature* **334**, 213–218.
32. Hamm,J., Kazmaier,M. and Mattaj,I.W. (1987) *EMBO J.* **6**, 3479–3485.
33. Patton,J.R., Patterson,R.J. and Pederson,T. (1987) *Mol. Cell. Biol.* **7**, 4030–4037.
34. Scherly,D., Boelens,W., van Venrooij,W.J., Dathan,N.A., Hamm,J., and Mattaj,I.W. (1989) *EMBO J.* **8**, 4163–4170.
35. Tschudi,C., Richards,F.F., and Ullu,E. (1986) *Nucl. Acids Res.* **14**, 8893–8903.
36. Gerke,V., and Steitz,J.A. (1986) *Cell* **47**, 973–984.
37. Tazi,J., Alibert,C., Temsamani,J., Reveillaud,I., Cathala,G. Brunel,C., and Jeanteur,P. (1986) *Cell* **47**, 755–766.
38. Zucker-Aprison,E., Thomas,J.D., and Blumenthal,T. (1988) *Nucl. Acids Res.* **16**, 7188.
39. Hashimoto,C., and Steitz,J.A. (1984) *Nucl. Acids Res.*, **12**, 3283–3293.

40. Krol,A., Lund,E. and Dahlberg,J.E. (1985) *EMBO J.* **4**, 1529–1535.
41. Ciliberto,G., Buckland,R., Cortese,R., and Philipson,L. (1985) *EMBO J.* **4**, 1537–1543.
42. Mattaj,I.W., and De Robertis,E.M. (1985) *Cell* **40**, 111–118.
43. Vankan,P. and Filipowicz,W. (1988) *EMBO J.* **7**, 791–799.
44. Vankan,P., Edoh,D. and Filipowicz,W. (1988) *Nucl. Acids Res.* **16**, 10425–10440.
45. Huang,X.-Y. and Hirsh,D. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 8640–8644.
46. Maroney,P.A., Hannon,G.J. and Nilsen,T.W. *Proc. Natl. Acad. Sci. USA*, **87**, 709–713.
47. Yuo,C.-Y., Ares,M. and Weiner,A.M. (1985) *Cell* **42**, 193–202.
48. Konings,D.A.M. and Mattaj,I.W. (1987) *Exp. Cell Res.* **172**, 329–339.
49. Shuster,E.O., and Guthrie,C. (1988) *Cell* **55**, 41–48.