PHYLOGENETIC COMPARISONS OF U2 SMALL NUCLEAR RNA SEQUENCES SUGGEST A PSEUDOKNOTTED STRUCTURE

Manuel Ares, Jr. and A. Haller Igel

Department of Biology, University of California
Santa Cruz, California 95064

ABSTRACT U2 small nuclear RNA (snRNA) is a highly conserved component of the cell nucleus, where it recognizes and binds to the intron branchpoint during splicing of premessenger RNA (premRNA). To develop an understanding of the secondary and higher order structure of U2 snRNA, we compared the sequences of U2 RNAs from ten organisms, aligned them, and analyzed the pattern of nucleotide changes. The highly conserved 5' half of the molecule contains two helices that may participate in a pseudoknot, adjacent to the nucleotides involved in pairing to the intron branchpoint. Although most U2 RNAs contain two 3' stem-loop structures (stems III and IV), both trypanosome U2 and a deleted but functional yeast U2 appear to lack stem III.

INTRODUCTION

Splicing of messenger RNA precursors requires a subset of small nuclear ribonucleoprotein particles (snRNPs) that are assembled with premRNA into a complex called the spliceosome (for reviews see 1-3). Early in spliceosome assembly, the U2 snRNP recognizes the intron branchpoint (4). In the yeast

1This work was supported by institutional start-up funds from the University of California, Santa Cruz.
Saccharomyces cerevisiae, part of the specificity of U2 snRNP-intron branchpoint recognition is contributed by base pairing between U2 snRNA and the intron, where appropriate U2 mutations suppress intron branchpoint mutations in a fashion consistent with Watson-Crick type RNA-RNA base pairing (5).

To further understand the mechanism of U2 action during splicing, we are studying the structure and function of U2 RNA. To generate phylogenetically consistent models of U2 RNA secondary structure, we compared U2 RNA sequences from divergent organisms. We found that the highly conserved 5' half of U2 contains the potential to form a pseudoknot, a recently recognized RNA secondary structure with some intriguing features (6). In the 3' half of the RNA, a region known as stem III seems not to be essential for function of both trypanosome and yeast U2 (E. Shuster and C. Guthrie, pers. comm.; M.A. and H.I., submitted). Because of the extreme conservation of the 47 residues at the 5' end (including the branchpoint recognition region), structure models for this part of the molecule cannot yet be generated with confidence.

RESULTS AND DISCUSSION

Comparative Analysis Reveals Eight Conserved Pairing Interactions

Alignment. The ability to identify paired regions by phylogenetic comparisons depends on the alignment of homologous sequences (7). The 5' 100-120 nucleotides (nt) of U2 RNA are highly conserved, and were easily aligned. The 3' half of the RNA is less well conserved, and varies in length from more than 1000 nt in yeast (8) to as few as 40 in trypanosome (9). Excluding these extreme cases, most U2 RNAs contain 80 nt that are readily aligned. We used human U2 as a standard sequence, and introduced stars into it to reserve space where insertions relative to human U2 occur in the other sequences. The other sequences were aligned with the human sequence, and stars were placed where
deletions relative to human U2 were apparent. A dash was placed where the sequence matched human U2. We did not attempt to include the entire yeast U2 sequence in our analysis; instead, we used the sequence of a 210 nt functional deletion derivative of yeast U2 constructed in our laboratory. We included trypanosome U2, even though it may not be functionally equivalent to the other U2 RNAs. The results of the alignment are shown in Figure 1.

**Positions 1-6.** U2 RNAs begin with an A residue blocked at the 5' end with a 2,2,7, trimethylguanosine cap. Positions 2-4 are conserved among the animals, but vary in bean, trypanosomes, and yeast, where there is also a single base insertion. Positions 5 and 6 are CU in all cases except yeast, where UC exists.

**Stem I.** Positions 7-26 have been proposed to form a stem-loop structure (10). There is phylogenetic support for the 12-14/19-21 interaction: all organisms have GCC/GGC here except worm and trypanosome, which have GCU/AGC. The lower part of the proposed stem would pair positions 7-9 with 24-26 (UCU/AGA). As yet, no variation in this sequence pairing has been found and it remains without phylogenetic support. In all organisms except worm and yeast, positions 10 and 11 are CG, while 22 and 23 are UA. The only difference in worm is an insertion of a U residue between positions 9 and 10, but this could affect the structure by pairing with A23. G11 could interact with U22, except that in yeast, all four positions are U. Perhaps canonical base pairs in this region are incompatible with function. The 15-18 loop consists of 4 U residues except for fly where position 17 is A, worm where position 16 is A, and trypanosome, where position 15 is A.

**The branchpoint recognition region.** The nucleotides from position 27 to 46 are the most highly conserved of U2, and include those that contact the intron branchpoint during splicing (5). Excluding trypanosomes, the only difference is the insertion of A between positions 30 and 31 in fly and worm. Guthrie and coworkers have shown that a G to U change at the yeast U2 equivalent of position 36 can suppress a UACUAAC to UAUAAC mutation, and
that a change equivalent to A35 to U suppresses a UACUAAC to UACAAAC mutation in a yeast actin intron (5). These experiments show that at least in yeast, positions 35 and 36 interact with the pre-mRNA branchpoint region during splicing, and suggest the formation of an extended helix including positions 33-38. Trypanosomes have 4-6 differences (depending on alignment) between all other U2 RNAs in this region. We chose to introduce a deletion of position 33 and an insertion of an A between residues 39 and 40, so that only one nucleotide within the 34-38 segment is changed, G36 to U, the same change as in one of the yeast U2 suppressors. Because there are few nucleotide changes elsewhere in the region, the possibility that these nucleotides interact among themselves or with other highly conserved sequences cannot be assessed.

The Pseudoknot. The nucleotides 47-52 are complementary to those at 61-65. This was first noted by Doug Black, however, because of high conservation of the sequence, the pairing was not supported by compensatory base changes. In addition, some support existed for pairing between 53-60 and 88-95, and formation of the two helices simultaneously was considered unlikely. A theoretical treatment of the possibility of

---

**FIGURE 1.** (Facing page) An alignment and comparison of U2 snRNAs from different organisms. The name of the organism is to the left of the line. Top, the 5' half of the RNA. Bottom, the 3' half. Sequences are from the following references: human, rat, mouse, chick, frog (Xenopus laevis), fly (Drosophila melanogaster), (11); worm, Caenorhabditis elegans, (J. Thomas and T. Blumenthal, personal communication); bean, Vicia faba, (12); Tryp., Trypanosoma brucei, (9); Yeast, Saccharomyces cerevisiae, (8). Boxed region connected by lines indicate pairings discussed in the text. Stars indicate nucleotide deletions, a dash indicates a match with human U2. The rounded box contains the Sm site, a sequence shared by anti-Sm antibody precipitable snRNPs (13), and which is essential for assembly of the RNA into RNP particles (14, 15). The position of this site in trypanosome U2 is not certain. Roman numerals refer to stem numbering used previously (10). Arrows above the yeast sequence indicate the possible extension of the stem IV pairing. Yeast 633-660 refers to a dispensable yeast U2 sequence with structure similar to stem III.
simultaneous formation and coaxial stacking of two such RNA helices was presented by Pleij and coworkers as part of an attempt to understand the recognition of 3' ends of plant RNA viruses by enzymes that recognize tRNA (6). The results of this treatment suggested that formation of such a structure, termed a pseudoknot, was possible.

With the addition of the yeast and worm sequences to the U2 database, phylogenetic support for both the 47-52/61-66 (labeled A in Figure 1) and the 53-60/88-95 helices (labeled B in Figure 1) appeared. Vertebrate, fly, and trypanosome U2 contain UAUCAG/CAGAUA in the A pairing, however in yeast the sequence reads UUUCAG/CAGAAA. Worm contains two sets of compensatory changes and a single change that converts an AU pair to GU: UAUCGU/ACGGUA. The pairing between 53-60 and 88-95 in vertebrates is UUUAUAUA/UAUAUAAA. This is supported by changes in yeast: UGUAACAA/UGUUAAGA.

Because of the nature of pseudoknots, all the base pairs described above cannot be formed if the helices are to exist simultaneously. This is represented in Figure 2 and is predicted by computer graphics modeling experiments (16), that suggest that at least one or two unpaired nucleotides would be needed to span the deep groove of the discontinuous A-helix. Therefore, position 53 in vertebrates, bean, and yeast is likely not to be paired in the pseudoknotted structure. This is also suggested by the presence of unpaired nucleotides at the equivalent positions of fly, worm and trypanosome U2. The 53/88 pair may be important if the region undergoes a conformational switch during U2 function.

The pairing between 68-73 and 79-84 (C, Figure 1), and the 74-78 loop are quite variable and well supported phylogenetically. 0-2 nucleotides separate the stem-loop from the pseudoknot on the 5' side, and to the 3' side is a region of variable length (2-13 nt) that connects the stem-loop to the helix formed by the B pairing (see Figures 1 and 2). Though technically not part of the pseudoknot, this stem-loop is part of the RNA strand that spans the shallow groove to complete the pseudoknot, and is probably stacked on the 47-52/61-66 helix.
Figure 2. Secondary structure models for different U2 RNAs. Top, human U2; Middle, trypanosome U2; Bottom, a functional deletion derivative of yeast U2. Inset, an alternative folding of yeast U2 nucleotides 1098-1120 to form a stem III-like structure. Numbering is not standardized to human U2 RNA.
The Sm site. Positions 99-105 are occupied by the Sm site, a sequence (consensus AU$_{3-6}$G) shared by many of the snRNAs (13). Trypanosome lacks a clear match to the consensus, however there is a 5/6 run of pyrimidines bounded by a purine on each side. Two purine residues (105 and 106) usually follow the Sm site, and thereafter a stretch of 4-5 nt that is highly variable, even among vertebrates.

Stem III. Positions 112-144 compose stem III. The bottom portion of the stem is a 7-8 base pair helix formed by pairing 112-120 with 137-144, that may (as in vertebrates) or may not (as in fly, worm, and bean) contain a single bulged residue. The top helix is made up of 124-127 paired to 132-135. This helix varies from 4 to 6 base pairs in length. The loop region and the closing base pair are invariant and are related to the tetranucleotide loop described by Tuerk et al. (17). The two helices are connected by a 3-5 nucleotide stretch (121-123) to the 5' side, and by a 0-2 nucleotide stretch 3' of the top helix.

Stem III may not be present in all functional U2 molecules. Trypanosome U2, which may or may not be functionally equivalent to other U2 RNAs, does not contain stem III. In addition, a functional deletion derivative of yeast U2 we have constructed (submitted for publication) appears to lack stem III, although as diagramed in Figure 2, the sequence can be folded to resemble either stem III or stem IV, but not both. We prefer the stem IV folding for two reasons: First, trypanosome U2 lacks stem III, and second, the stem III-like folding presented in the inset of Figure 2 contains an AU pair to close the loop, inconsistent with both the conservation of GC at this position in U2 (Figure 1), and the use of CG at the equivalent position of related tetranucleotide loops found in other RNAs (17). Although dispensable, the sequence from positions 633 to 660 of yeast U2 could form a convincing stem III (Figure 1).

Stem IV. At the 3' end of U2 RNA is another two-helix stem topped by a 10-13 nucleotide loop. The bottom helix of 6-8 base pairs is formed by pairing 147-152 with 179-184. The top helix of 4-6 base pairs forms between 154-158 and 172-176. Yeast
contains an unpaired A residue between 172 and 173, but the helix contains 8 pairs in this species. 0–3 unpaired residues can be found between the two helices to the 5' side of the top helix, while 2–4 unpaired residues separate them to the 3' side of the top helix.

The 159–171 loop is moderately well conserved. The most well conserved portion is the 163–167 region, which is GCAGU or GCACU. Trypanosomes have CCACU. To either side of these 5 nucleotides, there is greater variation, however it is not random. For example, yeast and trypanosomes are related, sharing 9 of 13 positions with each other, but only 5 of 13 with vertebrates.

The 3' end of U2 occurs 3–7 nucleotides past the base of stem IV, except in yeast. In yeast, stem IV can be extended an additional 9 base pairs (11 in our deletion derivative, see arrows, Figure 1), and the 3' end occurs 40 nucleotides past the end of this extension (22 in our deletion derivative).

ACKNOWLEDGEMENTS

We are grateful to Jeffrey Thomas and Tom Blumenthal for providing the worm U2 sequence prior to publication. We also thank Doug Black, Harry Noller, Kathy Parker, Ted Powers, Chris Tschudi and Elisabetta Ullu for stimulating discussions.

REFERENCES


