Extremely short 20–33 nucleotide introns are the standard length in *Paramecium tetraurelia*

Chris B. Russell*, Dean Fraga and Robert D. Hinrichsen
Basic Sciences Division, Fred Hutchinson Cancer Research Center, 1124 Columbia St, Seattle, WA 98104, USA

Received December 20, 1993; Revised and Accepted March 3, 1994

ABSTRACT

*Paramecium tetraurelia* has the shortest known introns as its standard intron length. Sequenced introns vary between 20 and 33 nucleotides in length. The intron sequences were discovered in genomic sequences coding for a variety of different proteins, including phosphatases, kinases, and low-molecular weight GTP-binding proteins. All intron sequences begin with the conserved dinucleotide GT and end with the conserved dinucleotide AG. The sequences are more AT rich than the *Paramecium* coding sequences. The identified sequences were confirmed as introns by sequencing several cDNA fragments. We report here analysis of the characteristics of 50 separate introns, including size, base composition, and a consensus sequence.

INTRODUCTION

The initial RNA generated by many eukaryotic genes consists of sequences (called exons) that are spliced together in the final product, while the intervening sequences (introns) are removed. The introns in protein-coding genes belong to the class of introns whose excision is catalyzed by a small nuclear RNA-protein complex called the spliceosome. These introns share certain common features that appear to be conserved in the eukaryotic world, and range in size from just over thirty nucleotides [in *Drosophila* up to thousands of nucleotides (1)].

*Paramecium tetraurelia* is a ciliated protozoan that diverged from the eukaryotic lineage leading to metazoans prior to the separation of plants, animals, and fungi (2). We have discovered that many *Paramecium tetraurelia* genes contain extremely short introns, between 20 and 33 nucleotides in length. These are the shortest identified introns known, and are the only size introns yet found in *Paramecium*. The size of the introns may have important implications for our understanding of the excision process.

MATERIALS AND METHODS

Strains and plasmids

*Paramecium* DNA was isolated from *Paramecium tetraurelia* strain 51s. PCR fragments were cloned into pBS vectors (Stratagene, La Jolla, USA) and transformed into *E. coli* XL-1 Blue or DH5-alpha.

PCR and sequencing

The introns examined were all found in clones derived from PCR amplifications from *Paramecium* DNA, using Perkin-Elmer Taq Polymerase. Degenerate primer sequences were designed to match conserved sequences in the serine–threonine protein phosphatases, protein kinases, or low-molecular-weight GTP-binding proteins. Clones were identified as the intended genes by homology with other known family members. The presence of the insert within *Paramecium* DNA was confirmed by Southern blots using the method of Church and Gilbert (3). *Paramecium* messenger RNA was prepared by extraction with guanidinium chloride (RNAzol, Molecular Research Center, Cincinnati), followed by isolation with oligo-dT columns (Stratagene). PCR of *Paramecium* cDNA was performed using the PCR–RNA kit from Perkin-Elmer. Control reactions which did not contain reverse transcriptase did not produce any visible product. PCR products were excised from SeaPlaque agarose (FMC) gels and purified with QIAEX beads (Qiagen).

Plasmid genomic clones were sequenced either with Sequenase (U.S. Biological) and [32P]dATP, or with dye primer sequencing (Applied Biosystems). Direct sequencing (without cloning) of genomic PCR products or cDNA PCR products was done by DyeDeoxy sequencing (Applied Biosystems).

Computer analysis

Computer searches for common motifs were made by word-search examination of the frequency and position of all possible nucleotide sequences of that size. The frequencies of the most common sequences were compared with the results for other sequences with the same nucleotide composition. Secondary structure was examined using the Mfold routine on the Wisconsin GCG program (4,5).

RESULTS

During the sequencing of DNA fragments generated by PCR of conserved gene segments in *Paramecium tetraurelia* it became obvious that the appropriate reading frame was not contiguous...
throughout many of the clones. The original clones were generated by PCR amplification of bulk genomic DNA with degenerate primers directed against conserved sequences either for the serine-threonine protein phosphatases (types 1, 2A, and 2B) (6,7) or for the low-molecular-weight GTP-binding proteins (8). The consensus amino acid sequences for these classes of proteins are highly conserved, and an appropriate reading frame was easily identified.

In several of these clones the reading frame is interrupted one or more times by 20–30 bp sequences. The added sequences are AT rich and always begin with the dinucleotide GT and end with the dinucleotide AG. We hypothesized that the extra sequences were in fact very small introns. Sequencing of PCR products amplified from cDNA confirmed this. We have also directly sequenced amplified portions of many genes, including some putative protein kinases, which show the presence of these small introns.

Fifty introns were analyzed. Figure 1 shows the sequence of some representative introns including the extremes in size; all the introns in Figure 1 were directly sequenced from uncloned genomic and cDNA amplifications. The second column in Figure 1 shows the nucleotide sequence determined from amplified cDNA. In all cases the putative intron sequence is cleanly removed. The majority of the other intron sequences in the data set are from PCR clones, and may contain rare artifacts as a result. The large number of sequenced introns reduce the effect such artifacts may have on the analysis, and the smaller set which have been sequenced directly would be free of such artifacts. Results from duplicate clones in the phosphatase PCR library indicate that the error level is about 1 in 2000.

A distinguishing characteristic of the ciliated protozoa is the separation of its genome into germline and somatic nuclei, with the somatic DNA being a processed form of the germline DNA. In Paramecium tetraurelia this processing involves cleavage of the germline sequences into smaller chromosomes, the excision of specific sequences called Internal Eliminated Sequences (9–11), and amplification to approximately 1000-fold in the somatic nucleus (see ref. 12 for a review). One alternative explanation concerning the additional sequences seen in Figure 1 is that rather than being introns, they are actually IES’s which are removed during development of the somatic nucleus. In Paramecium these sequences all have a T:A boundary on each end (11), which is very different than the intron sequences, with a GT and a AG on the 5’ and 3’ ends. One would also expect that the vast majority of clones would be derived from the somatic DNA, as that DNA is present at a few hundred-fold higher copy number than the germline DNA. We find that all of our DNA derived clones from the relevant genes do contain the intron sequences.

Table 1 shows the distribution of intron lengths. The smallest identified intron is 20bp in length, and the longest is 33 bp. The distribution of lengths appears centered about 25 bp, with an unexplained paucity of 24bp introns. The average length is 25.7 ± 2.5.

Tables 1 and 2 and Figure 3 show the relative frequency of each nucleotide at each position of the introns relative to the 5’s and 3’s ends. As can be seen, the introns all start with the

---

**Figure 1.** Sequences of some Paramecium tetraurelia introns, and adjacent exon sequences. The genomic DNA is illustrated in the left-hand column with the last eight bases of the 5’ exon in smaller letters, then the intron in larger letters between the two columns, followed by the first eight bases of the 3’ exon. The cDNA sequence in the right-hand column matches that of the genomic sequence with the intron removed. All sequences shown were determined from direct sequencing of uncloned PCR products. The consensus sequence is shown below (The symbol W stands for either A or T, and the symbol R stands for either A or G). Introns 1–5 are from the PP-l-omega-2 gene (6, GenBank accession numbers X67492, L26482), introns 6–8 are from two different low-molecular-weight GTP binding protein genes (8, GenBank accession numbers U03611-U03624), and introns 9 and 10 are from an uncharacterized putative protein kinase gene (GenBank accession numbers L26302–L26304). The entire data set is available on gopher as Computer and Network Resources/Software Archives/FHCRC FTP Server/pentinm.seq in gopher.fhcrc.org, and by anonymous ftp in the pub/seqdata directory on ftp.fhcrc.org.

---

**Figure 2.** The distribution of lengths in 50 introns of Paramecium tetraurelia. The average length is 25.7 ± 2.5 nucleotides. Genbank Accession numbers are L26269, L26297–L26301, L26481, L26482, L26302–L26304, U03611–U03625.

---

**Table 1.** Base composition of exons and introns at the 5’ splice site

<table>
<thead>
<tr>
<th>Position</th>
<th>-10</th>
<th>-5</th>
<th>-1</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>12</td>
<td>2</td>
<td>10</td>
<td>14</td>
<td>18</td>
<td>8</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>G</td>
<td>30</td>
<td>20</td>
<td>24</td>
<td>26</td>
<td>18</td>
<td>18</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>T</td>
<td>38</td>
<td>34</td>
<td>36</td>
<td>36</td>
<td>34</td>
<td>36</td>
<td>34</td>
<td>34</td>
</tr>
</tbody>
</table>

The nucleotide sequences of the introns aligned at the 5’ ends of the introns. The position -1 is the last nucleotide on the 5’ side of a splice junction and the position +1 is the first nucleotide on the 3’ side of the junction.
The consensus sequence for the 5' end is GTA(A/T)(A/T) and the consensus at the 3' end is (A/T)(A/T)TAG with a moderate frequency of (A/T)(A/T)AAG.

Tables 1 and 2 and Figure 3 also show summaries of the exon sequences just 5' and 3' of the introns. It should be noted that these introns are not randomly selected; the introns actually consist of sets of introns occupying the same location in genes coding for homologous proteins. As such there exists a non-random bias in the exon nucleotide frequencies. For example, the 19 phosphatase introns come from eight different locations, with not every gene having an intron at each location (some clones have no introns) (6). One location is represented five times. Even with such a non-random sample it appears there is a strong bias towards purine-rich sequences, and particularly a G residue at the first and last positions on each exon.

Analysis of all possible di-, tri-, tetra-, and pentanucleotide sequences did not yield any surprising anomalies in distribution or location within the introns. There are common sequences that may serve as acceptor sites, but there are no sequences that appeared at a drastically higher rate than other sequences. The sequence YRAT was common and the sequences TA and AT are found in all introns. Analysis of secondary structure yielded no clear pattern.

DISCUSSION

The short 20–33 nt length standard of Paramecium introns is unprecedented in the study of introns in other species (1). There exists one putative intron of 31 nucleotides in the Na⁺ channel

Table 2. Base composition of exons and introns at the 3' splice site

<table>
<thead>
<tr>
<th></th>
<th>-20</th>
<th>-15</th>
<th>-10</th>
<th>-5</th>
<th>-1</th>
<th>1</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>42</td>
<td>42</td>
<td>42</td>
<td>48</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>100</td>
</tr>
<tr>
<td>C</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>42</td>
<td>42</td>
<td>42</td>
<td>42</td>
<td>42</td>
<td>42</td>
<td>42</td>
<td></td>
</tr>
</tbody>
</table>

The nucleotide sequences of the introns aligned at the 3' ends of the introns. The position −1 is the last nucleotide on the 5' side of a splice junction and the position +1 is the first nucleotide on the 3' side of the junction.
gene of *Drosophila* (13,14), and another intron of 36 bp in the *Drosophila* vasa gene that has been sequenced in both genomic and cDNA clones (15). [A previous putative 3′nt intron in the White gene of *Drosophila*, cited in (1) has been revised (16)]. The bulk of *Drosophila* introns are above 50 nucleotides in length (1,17). The majority of the introns in the nematode *C.elegans* are 40–60 nt in length (18,19). It should be noted though that the methods used here would preclude the discovery of very large introns. While some of these genes have been examined on phage lambda genomic library clones without finding large introns, the original clones were all generated using PCR, under conditions where introns larger than 1000 nucleotides would have not been preferentially amplified. However, other genomic clones in the laboratory have failed to yield evidence of larger introns. Furthermore, nothing in the PCR techniques would have prevented amplification of moderate sized introns (100–300nt) had they been there.

The splicing site sequences of the short *Paramecium* introns are homologous to those of the larger introns found in other eukaryotes. The GT sequence at the 5′ splice site and the AG at the 3′ site are absolutely conserved. However, no clear branch point consensus sequence is observed. Yeast and metazoan introns are pyrimidine-rich near the 3′ splice site, while *Paramecium* introns have no apparent strong general bias, though some runs of T's are observed. Other sequences are observed within coding regions of *Paramecium* genes which would seem to satisfy the limited requirements for intron definition, but are not excised. Possibly some of the information for excision resides in areas away from the intron, or in the distance from previous introns. Some evidence from mammalian systems support the idea that distant information in exons may help define the cleavage sites (20).

*Paramecium*, along with *Tetrahymena*, belongs to the class of protists known as hlotrichous ciliates. One feature of the holotrichs is the presence of germline micronuclei containing standard chromosomes, and somatic macronuclei containing DNA which has undergone some processing and rearrangement (9,11). In hypotrichous ciliates this is carried to an extreme (10,22,23). In the processing of DNA in formation of the somatic macronuclei certain internal eliminated sequences (IES) are removed. We do not believe the sequences in this paper represent IES's. IES sequence data for *Paramecium* show a universally conserved T:A sequence at each boundary (11). These short *Paramecium* introns follow the GT...AG rule found in virtually all eukaryotic introns (1). Bulk DNA from *Paramecium* is predominantly somatic DNA, as the macronucleus is at an estimated ploidy of 1000 (12). PCR amplification from bulk DNA shows the presence of the introns, and does not yield product of the smaller size, while PCR from cDNA yields intronless sequences. *Paramecium* is related to the genus *Tetrahymena*, which contains introns of a larger size. Sequenced *Tetrahymena* introns range from 53 to 978 nucleotides long, with about half of them under 100 nucleotides (24). While all these organisms share several features such as separate germline micronuclei and somatic macronuclei, and an altered genetic code (23,25,26), several major aspects of these organisms have either diverged or been acquired by horizontal transmission during evolution. The short *Paramecium* introns are another phylogenetic distinction between these hlotrichous ciliates, which also includes differences in the number of germline nuclei (two in *Paramecium* vs. one in *Tetrahymena*) germline chromosome number, somatic nuclear ploidy, and exocytotic vesicle contents [see ref (27)]. This list of differences between the two organisms illustrates the phylogenetic distance between them. When one considers that some metazoa have different spliceosomes in different tissues [see ref (28) for example], a change in the spliceosomal mechanism is not a particularly large step in the context of the other differences between *Paramecium* and *Tetrahymena*, and very short introns may appear in other places in the eukaryotic world.

The distribution of intron frequencies is in itself interesting. Because most of these were originally discovered in PCR libraries of conserved sequence amplifications, the majority of the introns are contained in genes for which only partial clones exist at this time. The introns occur at a frequency of about one per 250 nucleotides of coding sequence in the kinase, phosphatase, and GTP-binding protein gene sequences examined. We have discovered no internal exon longer than 400 nucleotides. Other completely sequenced *Paramecium* genes include four surface antigen genes (29–33), calmodulin (34), a potassium channel (Tim Jegla, personal communication) and dihydrofolate reductase–thymidilate synthase (M.Schlichtherle and J.van Houten, personal communication), which have no introns, and a beta-tubulin with one intron of 27 nucleotides (35). There is also a hemoglobin gene in *Paramecium caudatum* with a single 23 nucleotide intron (36). Clearly the frequency of introns is highly variable in *Paramecium* genes.

While the splice site nucleotide sequences are clearly homologous with other eukaryotes, the short length of the *Paramecium* introns may not be sterically compatible with the splicing mechanism in eukaryotes that have only larger introns. This suggests that the physical structures of the spliceosomes may be qualitatively different, or that perhaps the *Paramecium* splicing process does not use the complete pathway of the other systems, perhaps not even forming a lariat.

Intron processing in other eukaryotes begins with the assembly of a spliceosome, where the U1 and U2 snRNPs bind to the 5′ splice site and the branch point, and the other snRNPs assemble on to this structure. [For a review of intron processing see ref. (37)]. The splice site recognition sequences of *Paramecium* introns are similar to those in other organisms; the 5′ end of each intron begins GU and the 3′ end has the sequence AG. Other nucleotides in the vicinity fit with patterns seen in other eukaryotes (26).

The first biochemical step in intron processing is cleavage of the 5′ splice site and formation of a branched lariat structure with the 5′ end of the intron being connected to the 2′ hydroxyl of a riboadenosine nucleotide. In *Saccharomyces cerevisiae* and metazoans there exist consensus sites for the lariat branchpoint which tends to be 18–36 nucleotides from the 3′ splice site (17,37). In *Paramecium* this size range instead approximates the size range of the full length of the introns. There is no obvious branch point consensus sequence in *Paramecium* introns. The longest sequences absolutely conserved in all the introns are the dinucleotides AT and TA. No degenerate pentanucleotides are universally found, nor is any one especially more favored than others. Several degenerate and non-degenerate heptanucleotide sequences were tested, including some resembling the mammalian consensus sites, and none were clearly more common than the others. The short size of the introns is perhaps sufficient to define a branch point as being any adenosine located within a range of a few nucleotides from one or the other end. Alternatively, the processing of these short introns may not require a branch
point. The short *Paramecium* intron may be small enough to allow splicing to be completed without interference from the free 5' end.

It might have been thought that the short length of *Paramecium* introns would be insufficient to allow the binding of the other snRNAs, as well as the geometric contortions necessary to form the lariat and splice junctions. While the snRNA genes are not yet identified in *Paramecium*, it appears that regular mRNA splicing does occur with introns as short as 20 nucleotides, and that the three dimensional structure of the intron and associated splicing components can accommodate all the necessary interactions. When viewed in light of the chemical similarities between intron splicing and RNA editing (38), one can imagine mechanisms that may take advantage of a universal short length for introns in a species. Further examination of the splicing system in *Paramecium* may discover a system intermediate between metazoan splicing and RNA editing, strengthening the analogy between them, and increasing our understanding of each.

ACKNOWLEDGEMENTS

This work was supported by NIH, NSF grants to RDH, NIH grant to DF, and an NIH core grant to FHCRC to support the BioComputing facility. Thanks to Mark Roth and Leo Kretzner for critical reading of the manuscript and to Mary Kay Dolejsi for automated DNA sequencing support.

REFERENCES