Nuclear pre-mRNA introns: analysis and comparison of intron sequences from *Tetrahymena thermophila* and other eukaryotes

Csilla Csank, Frances M.Taylor and Duane W.Martindale
Department of Microbiology, Macdonald College of McGill University, 21 111 Lakeshore Road, Ste Anne de Bellevue, PQ H9X 1C0, Canada

Received May 30, 1990; Revised and Accepted July 30, 1990

ABSTRACT

We have sequenced 14 introns from the ciliate *Tetrahymena thermophila* and include these in an analysis of the 27 intron sequences available from seven *T. thermophila* protein-encoding genes. Consensus 5′ and 3′ splice junctions were determined and found to resemble the junctions of other nuclear pre-mRNA introns. Unique features are noted and discussed. Overall the introns have a mean A+T content of 85% (21% higher than neighbouring exons) with smaller introns tending towards a higher A+T content. Approximately half of the introns are less than 100 bp. Introns from other organisms (approximately 30 for each) were also examined. The introns of *Dictyostelium discoideum*, *Caenorhabditis elegans* and *Drosophila melanogaster*, like those of *T. thermophila*, have a much higher mean A+T content than their neighbouring exons (>20%). Introns from plants, *Neurospora crassa* and *Schizosaccharomyces pombe* also have a significantly higher A+T content (10%-20%). Since a high A+T content is required for intron splicing in plants (58), the elevated A+T content in the introns of these other organisms may also be functionally significant. The introns of yeast (*Saccharomyces cerevisiae*) and mammals (humans) appear to lack this trait and thus in some aspects may be atypical. The polypyrimidine tract, so distinctive of vertebrate introns, is not a trait of the introns in the non-vertebrate organisms examined in this study.

INTRODUCTION

Pre-messenger RNA transcripts are processed to create mature translatable mRNA molecules. Crucial to this process is the precise excision of intervening sequences (introns) and ligation of exons. This event, pre-mRNA splicing (reviewed in 1, 2, 3), has been extensively studied in yeast (*Saccharomyces cerevisiae*) and mammalian systems where work has been facilitated by the development of *in vitro* splicing methodologies. Although the group I self splicing rRNA intron of the ciliated protozoan, *Tetrahymena thermophila*, has been intensively studied because of its unique catalytic abilities (reviewed in 4), little is known about nuclear mRNA introns in this or any other ciliate (5). Furthermore, despite the accumulation of intron sequence data from diverse groups of organisms, few comparative studies have been conducted.

Yeast (*S. cerevisiae*) and mammals share a common two-step mechanism for removal of nuclear mRNA introns (1, 2, 6). First, the bond that joins the 5′ exon to the intron is cleaved and an unusual 2′-5′ phosphodiester bond is made between the intron's 5′ G residue and an A residue (branch site) within the intron. Typically the branch site is located 18 to 40 bases upstream of the 3′ splice site (7, 8). In the second step, the bond joining the intron's 3′ terminal AG dinucleotide to the 3′ exon is split and the two exons are joined. The released intron is in the form of a lariat (6, 9).

The splicing reaction takes place in a multicomponent complex called a spliceosome (reviewed in 10). It is composed of splicing factors as well as the following four small nuclear ribonucleoproteins (snRNPs): U1, U2, U5, and U4 + U6 (11, 12). Although no catalytic function has been assigned to any of these components, those with an important role in the recognition of short stretches of conserved intron sequences have been identified. The U1 snRNA must base pair with conserved bases at the intron's 5′ splice junction for splicing to occur (13, 14). These bases are part of a 5′ intron consensus sequence that includes an invariant GT dinucleotide found at the 5′ intron border. The 3′ ends of nuclear pre-mRNA introns have a conserved terminal AG dinucleotide that is also part of a larger consensus sequence (1, 3, 8). This 3′ consensus sequence is thought to be recognized by a protein associated with the U5 snRNP (15, 16). Other intron sequences recognized by spliceosome components are less conserved between organisms. *S. cerevisiae* introns harbour an invariant branch site consensus sequence (TACTAAC) that base pairs with the U2 snRNA (17). Mammals have a less conserved branch site (7, 18, 19, 20) with which the U2 snRNA can base pair (21). More important in mammals than the branch site sequence itself is an adjacent downstream polypyrimidine tract (22, 72). A polypeptide required for the binding of the U2 snRNP to the branch site, is a polypyrimidine tract-recognizing factor (22, 23, 73). The polypyrimidine tract is usually found next to the 3′ splice site and hence has been included in the mammalian 3′ consensus sequence (24).

In this report we analyze 27 intron sequences from the ciliated

protozoan *T. thermophila* as well as representative introns from other eukaryotes. Features of the *T. thermophila* introns are compared with the introns of these other eukaryotes. Consensus sequences, as well as intron and exon sizes and nucleotide (nt) compositions, are explored. We show that a strong A+T enrichment in introns above levels in neighbouring exons is a trait common to introns from a diversity of eukaryotes including plants, a slime mold, a nematode, a ciliate, an insect, a filamentous fungus and a fission yeast, but is not a characteristic of a budding yeast (*S. cerevisiae*) or mammalian introns. The polypyrimidine tract, that is characteristic of mammalian introns is not a feature of the introns in the other organisms examined in this study.

MATERIALS AND METHODS

We determined the sequences of fourteen introns and flanking exons from two *T. thermophila* genes, ilsA (eight introns) and cnjB (six introns). The procedures used were as in (5). IlsA (formerly cupC) is an isoleucyl-tRNA synthetase gene (26) containing 8 introns, and cnjB is a conjugation-specific gene from which 12 introns have so far been sequenced (5; and this paper). The ilsA gene was first isolated as a partial cDNA (pC8; 26a). Complete genomic DNA and cDNA were obtained by screening a *T. thermophila* genomic DNA library (provided by K. Karrer, Brandeis Univ.) and a cDNA library (constructed for us by Stratagene from RNA isolated during early conjugation). The cnjB gene was isolated as described (5). A comparison of genomic and cDNA sequences confirmed the positions of introns from both these genes. Unpublished *T. thermophila* sequence data were kindly provided to us for two ribosomal protein genes, L21 (G. Rosendahl, P.H. Andraesen and K. Kristiansen, personal communication) and L1 (K. Kristiansen, H. Dresig, P.H. Andraesen, P. Hojrup, H. Nielsen and J. Engberg, personal communication). Other *T. thermophila* intron sequences were taken from the literature and include introns from the histone H1 gene (27), a calcium binding protein gene (28) and a third ribosomal protein gene, S25 (29). Sequence data for a nematode, *Caenorhabditis elegans* (30 introns; 30–36), and a slime mold, *Dictyostelium discoideum* (30 introns; 37–53), were from the literature. Sequence data for a filamentous fungus, *Neurospora crassa* (28 introns), a fission yeast, *Schizosaccharomyces pombe* (20 introns), an insect, *Drosophila melanogaster* (29 introns), two dicotyledonous plants, *Solanum tuberosum* (9 introns) and *Glycine max* (13 introns), a monocotyledonous plant, *Zea mays* (23 introns), and humans (39 introns) were obtained from version 59.0 or 63.0 of Genbank (Intelligenetics). The references to sequences for the budding yeast, *Saccharomyces cerevisiae*, were from Woolford (8) and most sequences were available through Genbank; they included 10 introns from ribosomal protein genes and 8 from other genes.

Computer-assisted DNA sequence analyses were done on an Apple Macintosh-plus computer using the DNA inspector IIe (Textco) and Pustell (IBI) programs.

Statistical analyses were done as in Sokal and Rohlf (54) using tables from Rohlf and Sokal (55). Kendall's nonparametric test or coefficient of rank correlation (τ) was used for testing associations between two variables. Student's T-test was used to test differences between two means.

RESULTS

The 5′ and 3′ ends of the 27 available *T. thermophila* introns, as well as the 3′ end of a partially sequenced *T. thermophila*

	5′exon ↓ intron		3′intron ↓ exon	size(nt)
1	CTTGATAAAA	GTATTAACATTAAAAAATGACTTTAAAATCTCATTTTAAT	TTTAAAATCTCATTTTAATTTCGAAATCTTAATATTTTAG ATATTTACTG	61
2	AGCTGAAGAG	GTACCTTATATAATTAATTCATTTATTCATTTATTGGTTT	GTTTTTGATGCAAGAGATTCATAATTTTTTTCTAAAATAG TAAAAAAAGA	76
3	CAGCTACTCT	GTTAAAGTAATTTTGGAATTGCCGCTATTTTATTTTTTAA	TATAAAAAAAATATAAAATTAATAAATTATTATCAAATAG TAATTTCGTT	94
4	ATTTCCCAAG	GTATTTATTTCTGAATAATTTATTATTTCTCAAAAGCTTC	ATTTTAAATATTAATGAATAATTCATTTTTTTTAAAATTAG AAGCAGAAAT	95
5	TGACATTTAG	GTAAAAAATAATTTTATTAATTTATCAATTAGACCATAAT	TATCAATTAGACCATAATTTATAAATTTATTATAAAATAG CAACACCTAA	62
6	ATCATAAAAA	GTATTATTTATTATCTTTTATTTACTTATTAAATAATATT	TTCATCATGACTTTATCATTAATTTTTTTTTTAAAAAAAG AGCCTCTAAG	364
7	AAAACTATTT	GTAATTAATTTAATTGAAATATTTTAGATAAAAATAATAT	AAATAATATTTTTAATGGTTTATAAAAAAATAACTATTAG ATATCCATAC	71
8	TATAAGCTAT	GTAATAAAATATTTATTTATTAAACAATCCATTCAATATT	TTTATACACAAGACAACTCTGATTTTGAATTATTTAATAG GGTCCATATC	676
9	TCAATTTAAG	GTAAAAATTAATCAATGAATTAATTAGTAAATTTAGAGTT	AATTAAATAAATAAATGATTTTTATTCATCAAAATATTAG AACTTATGCA	174
10	TCCTTTCTAG	GTATAAATTAGAATTTACTAATTCCCAATTTTTTTTACTT	GATTATTTTAATCTTCATCTAATCTATTTTAAATATTAAG AAAAAGGTTG	131
11	CCTTACAATC	GTAAATCTTAAATCAATCTATTACTTCATGTTTAAAAAAT	AATTATTTTTATGAGTTAAAATTTATAAAAATTAATAAAG AGTATTTCAT	211
12	GACTAATGAA	GTAATTAATTTATATTTTTATTTTTAACATACATACATAC	CATACATACCAACTTATTAAACACAATTTTCTTAATGAAG GTCAGACATT	71
13	TTCAAAGATC	GTAATAATTACTTCTACTAATAAATATGCCTTTTTCTATG	ATGTCCTAGAAATTAATATTTTCTTTTAATACTCATATAG TTTAGGAAAT	97
14	TACAATCAAG	GTATTTTATAAAAATTTTTATTTATTTCAGAAAATTCTAT	CTTAAATAAATAATATTAATATATTTTTAAATGCTTATAG GATGTCGTTT	225
15	AGTCTTAAAG	GTAATATTTAAGATAAAAATTAGCATTTTTAGTTAATTCG	GAAATACATTTAATCTAAAAATTATTTATTTTTATATTAG ATGGGTGTCG	300
16	GATTTGGTCG	GTAATGAAACTTCCTATTTAAATATCTTCTTTAATTAAAA	ATAAATTTAATTCAATATTTATTAATTAATTATATTATAG TTGGGTTGAC	192
17	GGCTTATTGG	GTAATTATTTAGATTGACATTTAAAAATATTATTTTACTT	AATATTATTTATACTTAAATTATTCCAAATATATTAATTAG GTTCCTAAAT	65
18	TTTATATATG	GTATTTATTTTCTTCTTTTAATAAAACCAAATTAATCATT	AATAAAACCAAATTAATCATTTTAATTCCTTATCCTAAAG ATCAACTCTC	59
19	ATGGATTATT	GTAAGAATATTAAAATAGCTATTTTTTAATTTTAAAACAA	AATAGCTATTTTTTAATTTTAAAACAACTTTTACAAATAG ATCTCCTGCC	53
20	ACCTATCATG	GTAATATTCAACAATTTATTATTATTATTAATGGTAATTA	TATTATTATTATTAATGGTAATTAATTTTTAAAATCAAAG TCTTTGACTG	56
21	AACGGTCAGC	GTAAGAATTTAACGATTTGTTTCTACTATTTTCATTAAAT	TTATTCATTTTATTGAAAATATCTTTAATTTAAAAATTAG GATGCCAAGA	354
22	AGGACCAAAG	GTGAAATATTTTTTTAATAACAATCAATCTACAGCAAGCA	TCTCAAAATAAATTTATTAAATTCAATTTTAAATAAAAAG ATGGGCTGAT	978
23	TCCAACACTG	GTAATTTGTTAATCTTGTTTTTTTTCTGATAAATGCTTTT	ATAATTTCAACGGCTTAATTAATATTTATTTTTAAAACAG GTAACTTCGG	197
24	ACCCCTCTAC	GTAAGCTATTAACACAAACTCAAAAAAACAATCTTTTAAA	ATTAAACATTAAAATAATCTTTCCTTTTAATTAATAATAG TGGTATTTTC	285
25	TCTAGAGTCG	GTAAGAGATCAATTACTTTTTAAATTGAATAAGTTTAAAT	TTTTTGATATAATTAACTAATCAATTAATTAAATTATTAG GTAACAACCA	908
26	TGCTGAAAAG	GTAATTAATTAAATTTTGAGTTAAATAAATATTAAACTAC	AGCTTGATTGATGTTAATTTTATTTTGAATTAATAAAAAG GTTGCTTTCG	682
27	GCTCGACGTC	GTAAGAAATAAATACATTTTATTCTTTTTTAGAACTTATA	TTAAAACGTAATTTTAAAATATCTTTTTAAATTAAAATAG GCTAGAAGAC	87
28			AAACATTTTTTCTTATAATCTAATTAATCTTTTCTAATAG GGTGCTAAGA	?

Fig. 1. Alignment of *T. thermophila* intron sequences. Shown are intron sequences closest to the 5′ splice site (40 nt) and 3′ splice site (40 nt) and flanking (10 nt) exon sequences. Underlined sequences are regions that overlap in the figure and are thus seen in both the 5′ and 3′ sequences. Dotted lines represent continuations of the sequence beyond regions shown (for introns longer than 80 nt). Introns are from the following genes: 1–12. cnjB, a conjugation-specific gene (4–9, ref. 5) 13–20. ilsA (formerly cupC), an isoleucyl-tRNA synthetase gene (CC and DM unpublished results; GenBank accession number: M30942); 21. a histone H1 gene (27); 22–26. three ribosomal protein genes; 22. S25 (29); 23–25. L21 (Rosendahl, G., Andraesen, P.H., Kristiansen, K., personal communication); and 26. L1 (Kristiansen, K., Dreisig, H., Andraesen, P.H., Hojrup, P., Nielsen, H., Engberg, J., personal communication); and 27–28. a calcium-binding protein gene (28).

intron, are aligned in Fig. 1. These alignments were used to formulate consensus sequences for the 5′ and 3′ intron junctions (Table 1). For the consensus sequences, a single nucleotide was included if its frequency at that position was greater than 40%. Two nucleotides were included if their frequency together was over 80% and that of each base over 30%. The *T. thermophila* intron consensus sequences are similar to those proposed by Martindale and Taylor (5) based on an examination of eight introns. Also in Table 1, the 5′ and 3′ consensus sequences for nuclear pre-mRNA introns from other organisms are presented. Human, *S. cerevisiae*, *S. pombe*, *C. elegans*, and *D. discoideum* consensus sequences were obtained by aligning and analyzing splice junction sequences of introns from these organisms (see Materials and Methods). The *S. cerevisiae* and *C. elegans* consensus sequences are extended to include additional base positions, but otherwise conform well to consensus sequences previously reported for these organisms (8, 57). To our knowledge thorough *D. discoideum* and *S. pombe* consensus sequences have not been published elsewhere. The nucleotide frequencies used in establishing the consensus sequences of vertebrate and plant introns are from Padgett et al. (3) and Brown (56) respectively.

We examined the introns of *T. thermophila*, *D. discoideum*, *C. elegans*, *S. pombe*, *S. cerevisiae*, and humans for preferences

Table 1. Intron junctions[a]

	5′ exon ↓ Intron									3′ Intron ↓ exon						
	-3	-2	-1	+1	+2	+3	+4	+5	+6...	...-5	-4	-3[b]	-2	-1		+1
vertebrate			A_{62}	G_{77} ↓	G_{100}	T_{100}	A_{60} G_{32}	A_{74}	G_{84} T_{50}	$(T/C)_n$	N	C_{78}	A_{100}	G_{100} ↓		G_{55}
human			C_{46} A_{51}	G_{79} ↓	G_{100}	T_{100}	A_{59} G_{38}	A_{69}	G_{95} T_{51}	$(T/C)_n$	N	C_{72}	A_{100}	G_{100} ↓		G_{41}
plant			A_{55}	G_{72} ↓	G_{100}	T_{99}	A_{70}	A_{55}	G_{65} T_{49}	T_n		G_{50}	C_{67}	A_{100}	G_{100} ↓	G_{60}
nematode *(C. elegans)*			A_{46} A_{73}	G_{70} ↓	G_{100}	T_{100}	A_{63}	A_{77}	G_{77} T_{53} T_{63} T_{57} $(A/T)_n$	$(A/T)_n$		T_{86}	T_{93}	T_{67}	C_{80} A_{100} G_{100} ↓	G_{60}
budding yeast *(S. cerevisiae)*			T_{44}	G_{50} ↓	G_{100}	T_{100}	A_{97}	T_{89}	G_{100} T_{94} T_{55} A_{33}	T_n		T_{61}	A_{44}	T_{55}	A_{100} G_{100} ↓ A_{33} T_{39} C_{44}	
fission yeast *(S.pombe)*			A_{45} A_{50}	G_{70} ↓	G_{100}	T_{100}	A_{95} A_{30}	T_{50}	G_{95} T_{70} T_n	T_{56}	N	T_{55}	N	T_{70}	A_{100} G_{100} ↓	
slime mold *(D. discoideum)*				↓	G_{100}	T_{100}	A_{93}	A_{70}	G_{76} T_{97} $(A/T)_n$	$(A/T)_n$		T_{63}	A_{70}	T_{70}	A_{100} G_{100} ↓	
ciliate *(T. thermophila)*		A_{52} T_{33}	A_{52} T_{33}	G_{56} ↓	G_{100}	T_{100}	A_{93}	A_{74}	T_{56} A_{56} T_{37} A_{59} T_{30} T_{48} A_{44} T_{78} $(A/T)_n$	$(A/T)_n$		A_{54}	A_{57} T_{39}	A_{67}	A_{64} T_{32}	T_{67} A_{100} G_{100} ↓

[a] Nucleotide frequencies (%) are subscripted.

[b] G residues are absent at this position in all introns surveyed except for a few exceptions in plant introns (56).

in nucleotide composition. Pyrimidine, purine, A+T, and G+C contents were analyzed at different nucleotide positions and graphed in a manner resembling that already done for plant and vertebrate introns (58, 59). Pyrimidine and purine frequencies of sequences surrounding the 5′ splice junctions and 3′ splice junctions are shown in Fig. 2. Similarly, the frequencies of A+T and G+C of the sequences surrounding the splice junctions are plotted against nucleotide position in Fig. 3.

T. thermophila, D. discoideum and *C. elegans* introns are similar in nucleotide composition. They show no preference for pyrimidines or purines (Fig. 2), but are A+T-rich (G+C-poor; Fig. 3) and thus resemble the introns of plants (58, 59). Excluding the invariant GT and AG dinucleotides, the average G+C content of these introns is 15% (range 6−24%) for *T. thermophila*, 8% (range 0.8−25%) for *D. discoideum*, and 24% (range 12−30%) for *C. elegans*. Their neighbouring exon sequences have a

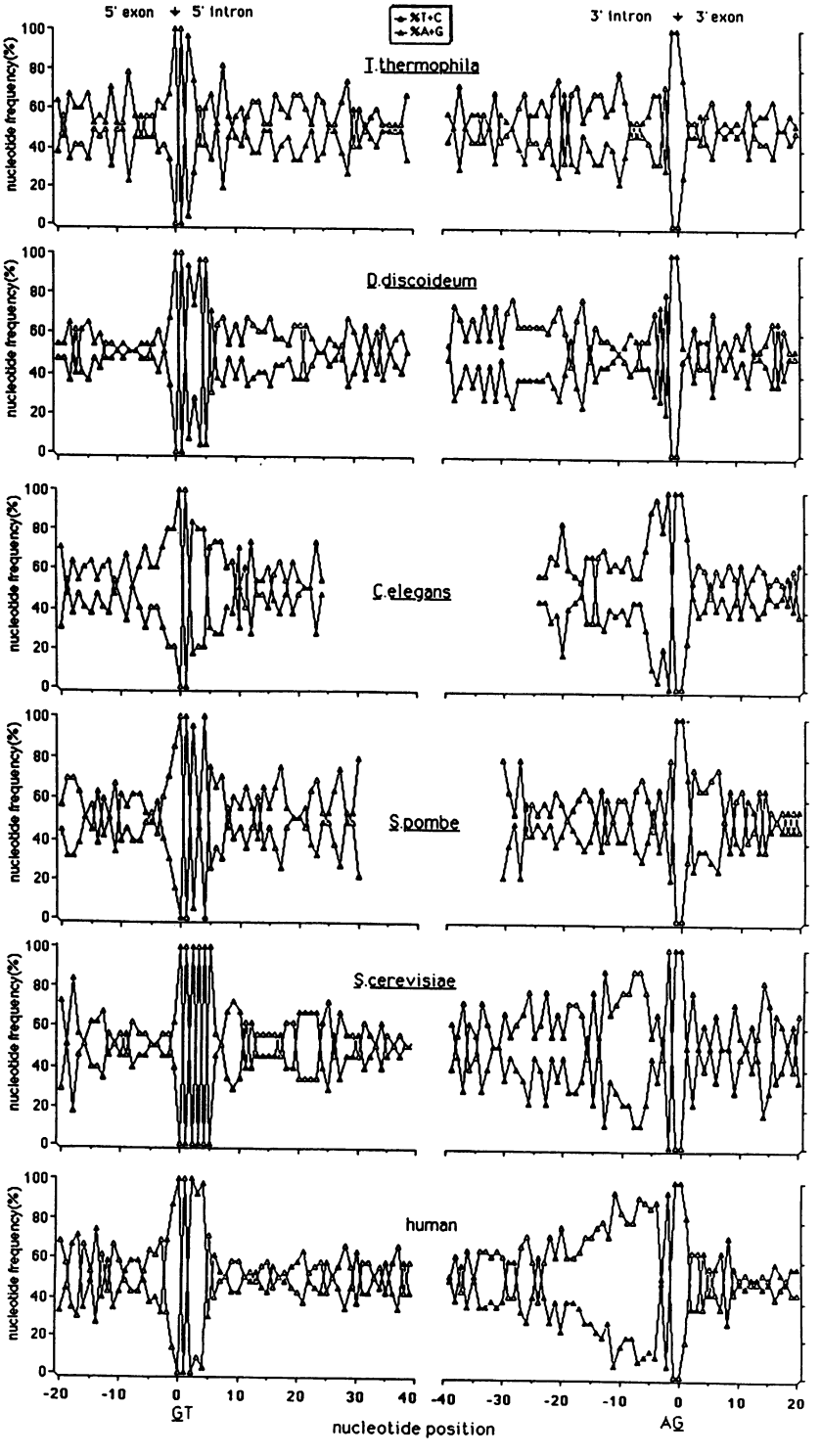

Fig. 2. Pyrimidine and purine composition of intron 5′ and 3′ sequences and flanking exons. The frequency of pyrimidines and purines are plotted against nucleotide position.

considerably higher G+C content as is evident in Fig. 3. The average coding exon G+C content is 41% (range 30−44%) for *T. thermophila*, 34% (range 27−41%) for *D. discoideum*, and 50% (range 41−59%) for *C. elegans*. The particularly high A+T content of intron sequences surrounding the 5′ and 3′ splice sites warranted the inclusion of A/T runs in the consensus sequences of *T. thermophila*, *C. elegans*, and *D. discoideum* introns (Table I). *S. pombe* introns are not enriched in purines or pyrimidines (Fig. 2) but are somewhat richer in A and T residues than their neighbouring exons (Fig. 3). This intron A+T enrichment is greater in the 30 nt preceding the 3′ splice site than in sequences following the 5′ splice site. However, there is a high frequency of T residues at the 5′ end of *S. pombe* introns. These are included as part of the consensus sequence (Table 1). The average G+C

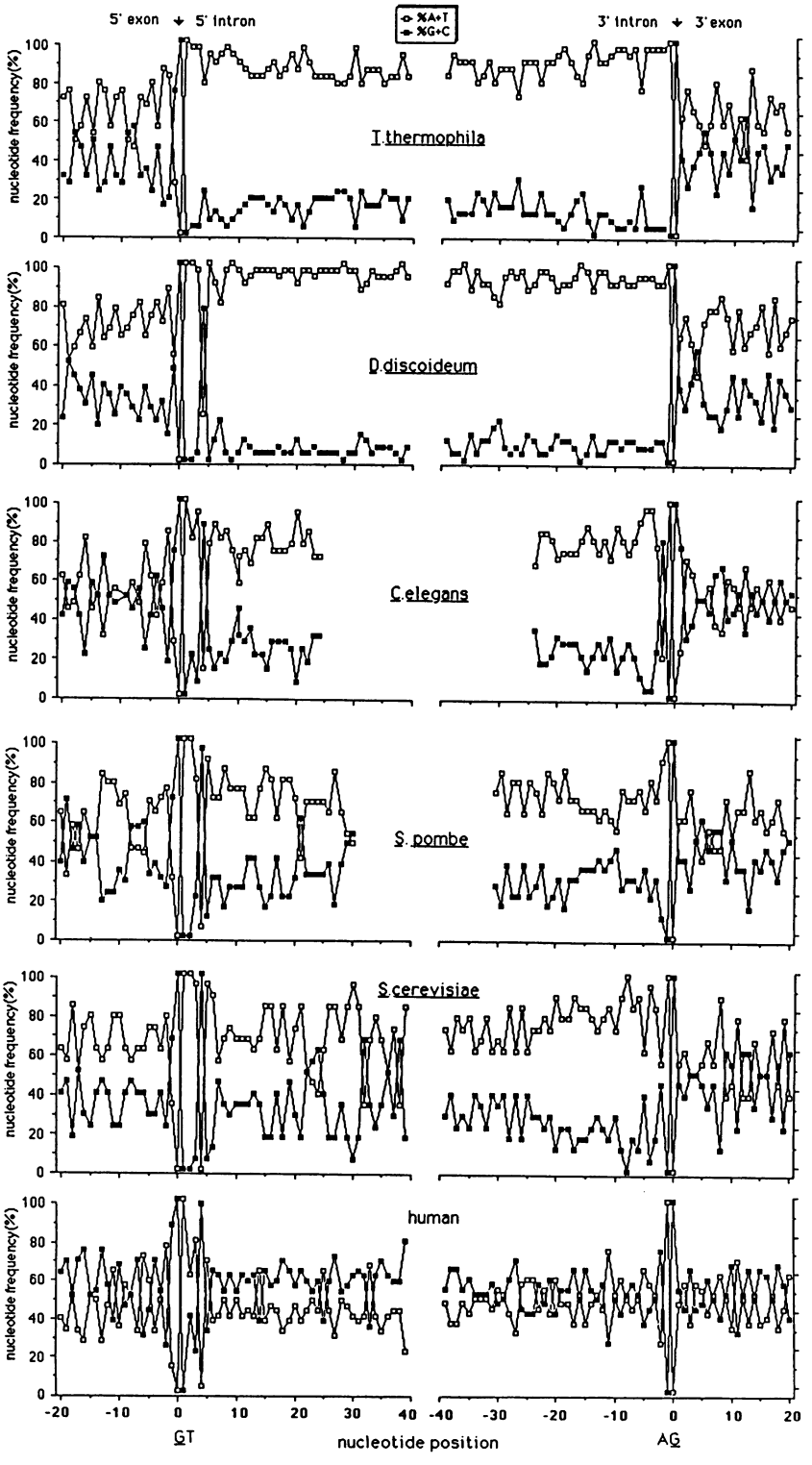

Fig. 3. A+T and G+C composition of intron 5′ and 3′ sequences and flanking exons. A+T and G+C frequencies are plotted against nucleotide position.

content of *S. pombe* introns is 29% (range 21−38%) and the average coding exon G+C content is 40% (range 37−43%)

The 5′ ends of *S. cerevisiae* and human introns show little or no difference in pyrimidine, purine, A+T, or G+C content relative to their neighbouring exon sequences (Fig. 2, 3). Near the 3′ end of *S. cerevisiae* introns, an increase in T residues (see Table 1) is responsible for the apparent rise in pyrimidine and A+T content. The numbers of the three other nucleotides decline in comparision to their 5′ end frequencies. A decrease is seen in the A+T content of the *S. cerevisiae* 3′ coding sequences in comparison to the coding regions flanking the 5′ splice site, but this may be an artifact of low sample number. Overall, *S. cerevisiae* introns have an average G+C content of 33% (range 21−37%) and their neighbouring exons an average of 41% (range 31−47%). Human intron 3′ sequences show a striking rise in pyrimidine content close to their 3′ splices sites (Fig. 2; Table 1), however neither 5′ nor 3′ intron sequences are A+T enriched (Fig. 3). The average G+C content of human introns is 54%, but the range is very wide (27−79%). The average G+C content of neighbouring exons is 56% with a smaller range (43−66%).

Since it became clear that, in many organisms, introns have a higher A+T content than their neighboring exons, we extended our study to include introns of an insect, *D. melanogaster*, a filamentous fungus, *N. crassa*, and three plants, *Z.mays*, *G. max* and *S. tuberosum*. The average% G+C of introns and average% G+C of neighboring coding exons were calculated for each organism. The difference between the two means is shown (Fig. 4). The mean intron and exon G+C contents were found to be significantly different at the 95% confidence level for all organisms but human. For *S. cerevisiae* the difference (8.5%) was small, but significant. When *S. cerevisiae* ribosomal protein genes were omitted from the analysis, the difference (4.8%) was insignificant. This discrepancy appears to be the result of a higher G+C content in the ribosomal protein gene exons and not a difference in intron G+C content.

The size distribution of *T. thermophila* introns is shown graphically in Fig. 5. About 50% of the *T. thermophila* introns are between 50 and 100 nt in length. The introns range in size from 53 to 978 nt, with a mean size of 241 nt. The longer *T. thermophila* introns are mainly from the ribosomal protein genes, a characteristic of introns from these types of genes even in organisms with predominantly small introns. We examined reports describing *D. discoideum* introns and found that small introns (below 150 nt) predominate in this organism (data not shown) as they do in fungi (60) (except *S. cerevisiae* (8)), insects (60) and *C. elegans* (57). Vertebrates and plants have fewer short introns and a wider intron size range (60).

The exons of *T. thermophila* analyzed in this study range in size from 61 to 886 nt (Fig. 6) and thus resemble the intron size range. The mean exon size of 330 nt is slightly larger than that of introns (241 nt) because fewer small exons exist. Although over 50% of exons are less than 250 nt long, only 15% are smaller than 100 nt. In other organisms there are also many exons smaller than 250 nt and few exons larger than 1000 nt (60).

In an analysis of fewer introns (eight), Martindale and Taylor (5), observed that small *T. thermophila* introns are more A+T rich than large introns. We examined our larger sample size to see if this trend was still evident. A positive correlation (τ = 0.51), significant at the 95% confidence level, was found between intron% G+C and intron size (Fig. 7). No significant correlation was found to exist between the G+C content of *T. thermophila* introns and neighboring flanking DNA, suggesting that intron G+C content is not affected by its surrounding DNA. The smaller *T. thermophila* introns (< 175 nt) are always less than 18% G+C (Fig. 7) as are the intron sequences (40 nt) adjacent to the splice sites of large introns. *D. discoideum* was the only other organism examined in this study that showed a positive correlation (τ = 0.35) significant at the 95% confidence level, between intron size and intron G+C content.

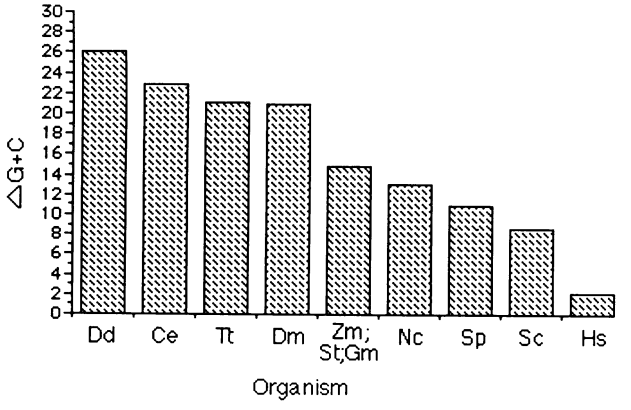

Fig. 4. Differences between the mean% G+C of exons and the mean% G+C of introns for different eukaryote species (% G+C exons -% G+C introns). Dd (*D. discoideum*); Ce (*C. elegans*); Tt (*T. thermophila*); Dm (*D. melanogaster*); Zm (*Z. mays*); St (*S. tuberosum*); Gm (*G. max*); Nc (*N. crassa*); Sp (*S. pombe*); Sc (*S. cerevisiae*); Hs (human). (For an extensive analysis of monocotyledonous and dicotyledonous plant introns refer to 58 and 59).

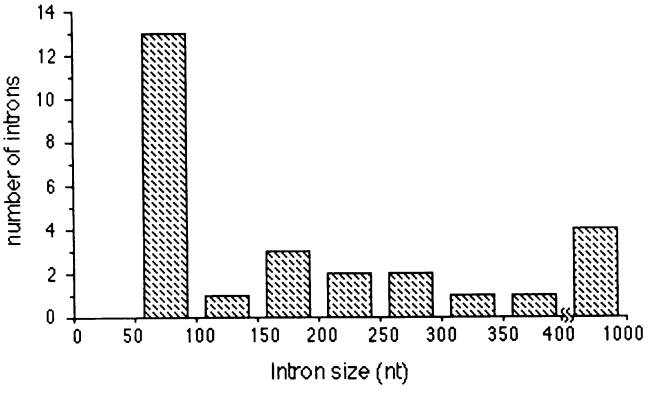

Fig. 5. *T. thermophila* intron size distribution.

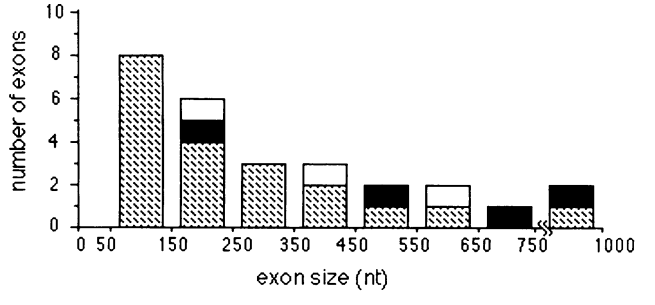

Fig. 6. *T. thermophila* exon size distribution. Internal exons are hatched boxes, 5′ terminal exons, open boxes, and 3′ terminal exons, black boxes.

DISCUSSION

We have determined the sequences of 20 *T. thermophila* introns (six were previously reported in 5). The sequences of these plus seven other *T. thermophila* introns (Fig. 1) were analyzed and compared to nuclear pre-mRNA introns from other organisms. Only *T. thermophila* introns were considered in this study since reports of introns from other ciliates are sparse. The sequences of two introns from *Tetrahymena pigmentosa* (61) have been deduced and are similar to the *T. thermophila* intron sequences. Two additional small *T. thermophila* introns have recently been reported from a histone H2A gene (74) and are similar to those analyzed in this paper. Out of a total of 16 known *T. thermophila* gene sequences, eight or half contain introns; five of these have multiple introns. This frequency is higher than previously estimated (5). It may be an overestimate because of the small sample size and the number of ribosomal protein genes (three) included since most ribosomal protein genes from lower eukaryotes have introns (8, 60). The proportion of genes without introns is higher in *T. thermophila* than in fungi (except *Saccharomyces* species), plants, insects, and vertebrates (60), and is consistent with the notion that rapidly-dividing lower eukaryotes have fewer introns than slowly-dividing higher organisms (62, 63).

The *T. thermophila* introns are always bounded by the invariant dinucleotides GT and AG that are part of the larger consensus sequences characteristic of nuclear pre-mRNA introns (Table 1; 3). Intron 5′ sequences, which in *S. cerevisiae* and mammals have been shown to associate with the U1 snRNA by base pairing (13, 14), are more conserved than intron 3′ sequences. *T. thermophila, S. cerevisiae, S. pombe*, and *D. discoideum* have a highly conserved A (>90%) at position +3 of the 5′ splice junction (Table 1), a characteristic of lower eukaryotes in general (64). A major difference between the *T. thermophila* 5′ consensus sequence and those of other organisms is found at positions +5 and +6 (Table 1). Instead of the usual G and T which dominate these positions respectively, *T. thermophila* prefers a T and an A. U1 snRNAs have been sequenced from a nematode (75), yeasts, plants, a fly, a frog, an algae, and vertebrates (10). The same eight contiguous nucleotides have been found at the 5′ end of all U1 snRNAs examined and these complement the prototypical 5′ consensus sequence AG/GTAAGT (64). The variability in 5′ consensus sequences observed within and between species (Table 1; 64) contrasts with this perfect conservation of U1 snRNA bases theoretically proposed to base pair with the 5′ splice site (64). A perfect match between the 5′ splice site bases and those of U1 snRNA does not seem to be important for splicing in yeast since U1 base pairs at a few but not all 5′ splice site positions (14, 76). Indeed, the geometry of association between U1 snRNAs and 5′ splice sites seems to be flexible with stable interactions occurring with a diversity of splice site sequences (64). The actual mechanisms that result in selection and cleavage of the 5′ splice site are as yet poorly understood (14, 64).

Fig. 7. Scatter plot of intron G+C content versus intron size.

Sequences upstream of the 3′ ends of *T. thermophila, D. discoideum, C. elegans, S. pombe, S. cerevisiae* (Fig. 2) and plant (58, 59) introns are not highly enriched in pyrimidines. The 3′ end of human (Table 1; Fig. 2) and all vertebrate introns (Table 1, 3, 59), however, is characteristically preceded by a stretch of pyrimidines. In mammals, this polypyrimidine tract is required for branching and spliceosome assembly and appears to be recognized by a polypeptide that binds before the U2 snRNP can associate with the branch site (22, 23, 73). In contrast, polypyrimidines are not needed at the 3′ ends of *S. cerevisiae* introns (8, 65), at least not for branching and spliceosome formation (77); nor are pyrimidine runs required in plant introns (58). Despite this, the T-content (but not C-content) of sequences immediately upstream of the 3′ ends of yeast and plant introns is high enough to be part of their 3′ consensus sequences (Table 1). It has been suggested (78) that T-runs may be of importance in the splicing of fungal introns when branch sites are at a distance from the 3′ splice site, as in *S. cerevisiae* (8), and not when branch sites are next to their 3′ splice site, as in *S. pombe* (79). Indeed, the frequency of introns with T-runs (>4nt) within the 30 nt upstream of their 3′ ends is much higher in *S. cerevisiae* (72%) than in *S. pombe* (45%). In fact, *S. pombe* is the only lower eukaryote examined (Fig. 2; 3) that has fewer T-runs at the 3′ end of its introns than at the 5′ end. In general, T-runs occur about 20% more often at intron 3′ ends than at intron 5′ ends. In all organisms examined, but *S. pombe*, greater than 60% of the introns have at least one T-run in the sequences 25 nt upstream of their 3′ splice sites. The length of these T-runs reaches a maximum of 13 nt in *S. cerevisiae*, 10 nt in *T. thermophila*, and in *D. discoideum* long homopolymers (>30 nt) of A or T residues occur at 3′ as well as at 5′ ends. Therefore, although most organisms lack the accentuated polypyrimidine tracts of vertebrates, short T-runs do occur preferentially upstream of 3′ splice sites in many organisms. These T-runs could help splicing factors locate 3′ splice sites, or may play a non-specific role in limiting secondary structure or reducing the likelihood of AG dinucleotides occurring in this region.

Throughout their intron sequences, including intron boundaries, *T. thermophila, C. elegans*, and *D. discoideum* have a very high A+T content (Fig. 3). A/T runs are included as part of their consensus sequences because at all positions analyzed (Fig.3) either an A residue is present greater than 40% of the time or a T residue is present greater than 40% of the time. The analyses presented here suggest that the introns of most organisms have a higher A+T content than their neighboring exons (Fig. 4) and extend previous observations that the introns of plants, *C. elegans, T. thermophila, D. discoideum* and chloroplasts (not dealt with here) are all A+T-rich (58, 59). This trait may however have been lost from human and *S. cerevisiae* introns during evolution (Fig. 3, 4). The difference between mean exon% G+C and mean intron% G+C is most striking (>20%) for the introns of *D. discoideum, C. elegans, T. thermophila* and *D. melanogaster*. The introns of plants (also see 58), the filamentous fungus, *N. crassa*, and the fission yeast, *S. pombe*, are also richer (10–20%)

in A and T residues than their neighboring exons (Fig. 4). When tested statistically the difference between intron and exon G+C content had no relationship to the G+C content of an organism's coding regions.

Recently, splicing determinants of plant intron sequences have been deduced using plant protoplasts (58). A high A+T content (higher than in surrounding exons) and appropriate splice site consensus sequences are necessary for the splicing of plant introns. This requirement for A and T residues in plant introns is independent of position. Since splicing can be restored in synthetic G+C-rich introns by inserting A and T residues, a positive role for A+T-richness was suggested (58). This is the first direct evidence that A+T-richness may be needed for some introns to be removed.

Since plant introns require A+T-richness for splicing, it is probable that other A+T-rich eukaryote introns do also, especially those of nematodes, slime molds, insects, and ciliates. The observation that small *T. thermophila* and *D. discoideum* introns have a tendency to be more A+T-rich than larger introns (see Results) also suggests that a high A+T content especially around the splice sites of these organisms is required, since a large proportion of a small intron is likely to be crucial for splicing. Thus, it is probable that intron A+T-richness is not just the result of a random drift towards a higher A+T content in a genome's non-coding DNA. The high A+T content of the introns of several organisms and the frequent presence of A and T homopolymers could limit the formation of strong secondary structure in introns. Secondary structure analysis of plant introns has shown that their RNA sequences are largely single-stranded (58). It has also been demonstrated that the insertion of inverted repeats between the branch point and 3′ splice site of yeast (66) and mammalian (22) introns interferes with splicing, suggesting a need to limit secondary structure in this region. A lack of strong secondary structure in general may help splicing factors access binding sites on intron sequences. Another possibility is that A+T-rich sequences serve specifically as binding sites for factors important for splicing (58) as suggested for the mammalian polypyrimidine tract (22, 73). A+T-rich intron sequences may also help splicing factors differentiate between exons and introns. In vertebrate genes with multiple introns, there is evidence that the three snRNPs (U1, U2 and U5) interact in a concerted manner to recognize both ends of an exon in a process termed exon definition (80). An appropriate 5′ splice site near an upstream 3′ splice site is chosen (80). The splicing factors of organisms with A+T-rich introns could search downstream exon sequences and choose a 5′ splice site at the start of an A+T-rich region. The exon definition hypothesis also suggests that internal exons (those without poly (A) or cap sites) should be short (80). Supporting this we find that *T. thermophila* internal exons are mostly less than 250 nt; 5′ and 3′ terminal exons are in general longer (Fig. 6). A similar trend is observed in vertebrates (80).

Even though branching and lariat formation have not yet been demonstrated experimentally for cis-splicing systems other than yeast (9) and mammals (6), the branch site recognizing snRNP (U2) has been identified in *D. melanogaster, C. elegans,* a trypanosome, a bird, an amphibian, and a plant (10). This suggests that branching and lariat formation are universal features of nuclear pre-mRNA splicing. The 3′ splice site in mammals (22) and probably yeast (67) appears to be determined by a process that scans the sequence downstream of the branch site for the first AG dinucleotide. Avoidance of AG between the branch site and 3′ splice site is therefore critical. Gelfand (24) noted the absence of AG dinucleotides in the 25 nucleotides upstream of the 3′ splice site of mammalian introns. We find that the *T. thermophila* introns similarly lack AG dinucleotides within the 24 nucleotides upstream of the 3′ splice sites. In fact, G residues are very rare in the 40 nt upstream of the 3′ splice site. Furthermore, in all organisms examined, a pyrimidine precedes the AG dinucleotide of the 3′ splice site at position −3, and G residues are absent with rare exceptions in plants (Table 1; 3, 56). Indeed, in mammals GAG trinucleotides are not cleaved efficiently although they are accurately recognized by the splicing machinery (22). We have also examined *T. thermophila* introns for the mammalian-like branch point sequence, PyTPuAPy (24), downstream of internal AG dinucleotides. The only conserved sequence that fits the consensus is TTAAT, found within the 40 nt upstream of the 3′ end in 65% of the introns. However, this sequence is present almost as frequently at the 5′ ends of introns. No other conserved sequences are evident. Selection of a branch point in *T. thermophila* introns may depend on elements other than a conserved branch point sequence. In support of this is the finding that the branch site of trans-spliced introns of trypanosomes occurs at an A residue, but there are no other conserved features in the sequences surrounding this site (68). Additionally, specific sequences do not seem to be needed for branch formation in plants (58), although mammalian-like branch site sequences are often located upstream of their intron 3′ splice junctions (56) and do promote most efficient splicing (58).

Evidence for the existence of different splicing mechanisms in various organisms is found from studies of intron splicing in heterologous systems. *S. cerevisiae* introns can be accurately removed in human cell extracts (69) but not vice-versa (70), although mammalian introns can be removed in *S. pombe* cells (81). Plant introns can usually be removed in human nuclear extracts, whereas human introns are not excised accurately in plant cells (59, 71). Similar studies with the introns from a diversity of organisms in a variety of heterologous systems may be revealing.

In this report, we have examined the mRNA intron sequences of a number of eukaryotic species. These include a ciliated protozoan, a slime mold, a nematode, an insect, plants, a filamentous fungus, a fission yeast, a budding yeast and a vertebrate (humans). Except for those of the budding yeast (*S. cerevisiae*) and humans, the introns examined were found to be A+T-rich. In addition, we observe that the polypyrimidine tract, present at the 3′ end of vertebrate introns (59), is not present in the introns of the non-vertebrate species tested, although short T-runs are common. The introns of mammals and budding yeast may in some ways be atypical, which emphasizes the importance of examining splicing requirements in a diversity of organisms. Such studies may give us clues to how the splicing of pre-mRNA molecules evolved.

ACKNOWLEDGEMENTS

CC was supported by a Postgraduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC), a McConnell Fellowship from McGill University, and a graduate fellowship from the DesJardins Foundation (Bourse Girardin-Vaillancourt). FMT was supported by a graduate fellowship from the government of Quebec (Fonds pour la Formation de Chercheurs et l'aide a la recherche). This work was supported by an Operating Grant to DM from NSERC. We thank Marlene Parkinson for her expert typing.

REFERENCES

1. Green, M.R. (1986) *Ann. Rev. Genet.*, **20**, 671–708.
2. Sharp, P.A. (1987) *Science*, **235**, 766–771.
3. Padgett, R.A., Grabowski, P.J., Konarska, M.M., Seiler, S., Sharp, P.A. (1986) *Ann. Rev. Biochem.*, **55**, 1119–1150.
4. Cech, T.R., Bass, B.L. (1986) *Ann. Rev. Biochem.*, **56**, 599–629.
5. Martindale, D.W., Taylor, F.M. (1988) *Nucleic Acids Res.*, **16**, 2189–2199.
6. Konarska, M.M., Grabowski, P.J., Padgett, R.A., Sharp, P.A. (1985) *Nature*, **313**, 552–557.
7. Reed, R., Maniatis, T. (1985) *Cell*, **41**, 95–105.
8. Woolford, J.L. (1989) *Yeast*, **5**, 439–457.
9. Domdey, H., Apostol, B., Lin, R.J., Newman, A., Brody, E., Abelson, J. (1984) *Cell*, **39**, 611–621.
10. Guthrie, C., Patterson, B. (1988) *Ann. Rev. Genet.*, **22**, 387–419.
11. Reed, R., Griffith, J., Maniatis, T. (1988) *Cell*, **53**, 949–961.
12. Fu, X.D., Maniatis, T. (1990) *Nature*, **343**, 437–441.
13. Zhuang, Y., Weiner, A.M. (1986) *Cell*, **46**. 827–835.
14. Siliciano, P.G., Guthrie, C. (1988) *Genes Dev.*, **2**, 1258–1267.
15. Gerke, V., Steitz, J.A. (1986) *Cell*, **47**, 973–984.
16. Tazi, J., Alibert, C., Temsamani, J., Reveillaud, J., Cathala, G., Brunel, C., Janteur, P. (1986) *Cell*, **47**, 755–766.
17. Parker, R., Siliciano, P.G., Guthrie, C. (1987) *Cell*, **49**, 229–239.
18. Zhuang, Y., Goldstein, A.M., Weiner, A.M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 2752–2756.
19. Black, D.L., Chabot, B., Steitz, J.A. (1985) *Cell*, **42**, 737–750.
20. Reed, R., Maniatis, T. (1988) *Genes Dev.*, **2**, 1268–1276.
21. Wu, J., Manley, J.L. (1989) *Genes Dev.*, **3**, 1553–1561.
22. Smith, C.J.W., Porro, E.B., Patton, J.G., Nadal-Ginard, B. (1989) *Nature*, **342**, 243–247.
23. Ruskin, B., Zamore, P.D., Green, M.R. (1988) *Cell*, **52**, 207–219.
24. Gelfand, M.S. (1989) *Nucleic Acids Res.*, **17**, 6369–6382.
25. Martindale, D.W., Martindale, H.M., Bruns, P.J. (1986) *Nucleic Acids Res.*, **14**, 1341–1353.
26. Martindale, D.W., Gu, Z.M., Csank, C. (1989) *Curr. Genet.*, **15**, 99–106.

26a. Martindale, D.W., Bruns, P.J. (1983) *Mol. Cell. Biol.*, **3**, 1857–1865.

27. Wu, M., Allis, C.D., Richman, R., Cook, R.G., Gorovsky, M.A. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 8674–8678.
28. Takamesa, T., Ohnishi, K., Kobayashi, T., Takagi, T., Konishi, K., Watanabe, Y. (1989) *J. Biol. Chem.*, **264**, 19293–19301.
29. Nielsen, H., Andraesen, P.H., Dreisig, H., Kristansen, K., Engberg, J. (1986) *EMBO*, **5**, 2711–2717.
30. Heschl, M.F.P., Baillie, D.L. (1989) *DNA*, **8**, 233–243.
31. Spieth, J., Denison, K., Zucker, E., Blumenthal, T. (1985) *Nucleic Acids Res.*, **13**, 7129–7137.
32. Karn, J., Brenner, S., Barnett, L. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 4253–4257.
33. Kramer, J.M., Cox, G.N., Hirsch, D. (1982) *Cell*, **30**, 599–606.
34. Greenwald, I. (1985) *Cell*, **43**, 583–590.
35. Yarbough, P.O., Hayden, M.A., Dunn, L.A., Vermersch, P.S., Klass, M.R., Hecht, R.M. (1987) *Biochim. Biophys. Acta.*, **908**, 21–33.
36. Cox, G.N., Fields, C., Kramer, J.M., Rosenzweig, B., Hirsch, D. (1989) *Gene*, **76**, 331–334.
37. Steel, L.F., Smyth, A., Jacobson, A. (1987) *Nucleic Acids Res.*, **15**, 10285–10298.
38. Reymond, C.D., Gomer, R.H., Mehdy, M.C., Firtel, R.A. (1984) *Cell*, **39**, 141–148.
39. Podgorski, G.L., Franke, J., Faure, M., Kessin, R.H. (1989) *Mol. Cell. Biol.*, **9**, 3938–3950.
40. Fosnaugh, K.L., Loomis, W.F. (1989) *Mol. Cell. Biol.* **9**, 5215–5218.
41. Muller-Taubenburger, A., Westphal, M., Noegel, A., Gerisch, G. (1989) *FEB*, **246**, 185–192.
42. Fosnaugh, K.L., Loomis, W.F. (1989) *Nucleic Acids Res.*, **17**, 9489.
43. Witke, W., Noegel, A.A. (1990) *J. Biol. Chem.*, **265**, 34–39.
44. Noegel, A., Witke, W., Schleicher, M. (1987) *FEB*, **221**, 391–396.
45. Hopkinson, S.B., Pollenz, R.S., Drummond, I., Chisholm, R.L. (1989) *Mol. Cell. Biol.*, **9**, 4170–4178.
46. Dynes, J.L., Firtel, R.A. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 7966–7970.
47. Loomis, W.F., Fuller, D.L. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 886–890.
48. Singleton, C.K., Manning, S.S., Ken, R. (1989) *Nucleic Acids Res.*, **17**, 9679–9692.
49. Ayres, K., Neuman, W., Rowekamp, W.G., Chung, S. (1987) *Mol. Cell. Biol.*, **7**, 1823–1829.
50. Datta, S., Firtel, R.A. (1987) *Mol. Cell. Biol.*, **7**, 149–159.
51. Mann, S.K.O., Firtel, R.A. (1987) *Mol. Cell. Biol.*, **7**, 458–469.
52. Kimmel, A.R., Firtel, R.A. (1980) *Nucleic Acids Res.*, **8**, 5599–5610.
53. Ragheb, J.A., Dottin, R.P. (1987) *Nucleic Acids Res.*, **9**, 3891–3906.
54. Sokal, R.R., Rohlf, F.J. (1981) Biometry. W.H. Freeman and Co., San Francisco.
55. Rohlf, F.J., Sokal, R.R. (1981) Statistical Tables. W.H. Freeman and Co., New York.
56. Brown, J.W.S. (1986) *Nucleic Acids Res.*, **24**, 9549–9559.
57. Blumenthal, T., Thomas, J. (1988) *TIG*, **4**, 305–308.
58. Goodall, G.J., Filipowicz, W. (1989) *Cell*, **58**, 473–483.
59. Wiebauer, K., Herrero, J.J., Filipowicz, W. (1988) *Mol. Cell. Biol.*, **8**, 2042–2051.
60. Engbert, J., Bojsen, K., Nielsen, H. (1989) in Cech, T.R. (ed.), Molecular Biology of RNA. Alan R. Liss, Inc., pp. 145–154.
61. Hawkins, J.D. (1986) *Nucleic Acids Res.*, **16**, 9893–9908.
62. Gilbert, W., Marchionni, M., McKnight, G. (1986) *Cell*, **46**, 151–154.
63. Fink, G.R. (1987) *Cell*, **49**, 5–6.
64. Jacob, M., Gallinaro, H. (1989) *Nucleic Acids Res.*, **17**, 2169–2180.
65. Fouser, L.A., Friesen, J.D. (1987) *Mol. Cell. Biol.*, **7**, 225–230.
66. Halfter, H., Gallwitz, D. (1988) *Nucleic Acids Res.*, **16**, 10413–10423.
67. Patzelt, E., Perry, K.L., Agabian, N. (1989) *Mol. Cell. Biol.* **9**, 4291–4297.
68. Langford, C.J., Klinz, F.J., Donath, C., Gallwitz, D. (1984) *Cell*, **36**, 645–653.
69. Ruskin, B., Pikielny, C.W., Rosbash, M., Green, M.R. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 2022–2026.
70. Langford, C., Nellen, W., Niessing, J., Gallwitz, D. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1496–1500.
71. VanSanten, V.L., Spritz, R.A. (1987) Gene, **56**, 253–265.
72. Reed, R. (1989) Genes Dev., **3**, 2113–2123.
73. Garcia-Blanco, M.A., Jamison, S.F., Sharp, P.A. (1989) Genes Dev., **3**, 1984–1986.
74. van Daal, A., White, E.M., Elgin, S.C.R., Gorovsky, M.A. (1990) J. Mol Evol., **30**, 449–455.
75. Thomas, J., Lea, K., Kucker-Aprison, E., Blumenthal, T. (1990) Nucleic Acids Res., **18**, 2633–2642.
76. Séraphin, B., Rosbash, M. (1989) Gene, **82**, 145–151.
77. Rymond, B.C., Torrey, D.D., Rosbash, M. (1987) Genes Dev., **1**, 238–246.
78. Parker, R., Patterson, B. (1987) in Inouye, M., Dudock, B.S. (ed.), Molecular Biology of RNA. Academic Press, Inc., pp. 133–149.
79. Hindley, J., Phear, G., Stein, M., Beach, D. (1987) Mol. Cell. Biol., **7**, 504–511.
80. Robberson, B.L., Cote, C.J., Berget, S.M. (1990). Mol. Cell. Biol., **10**, 84–94.
81. Kaufer, N.F., Simanis, V., Nurse, P. (1985) Nature, **318**, 78–80.