Volume 14 Number 24 1986 Nucleic Acids Research

A catalogue of splice junction and putative branch point sequences from plant introns

John W.S.Brown

Institute for Biology III, Albert Ludwigs University, D-7800 Freiburg, FRG

Received 10 October 1986; Accepted 14 November 1986

ABSTRACT

Splice junction and possible branch point sequences have been collected from 177 plant introns. Consensus sequences for the 5` and 3` splice junctions and for possible branch points have been derived. The splice junction consensus sequences were virtually identical to those of animal introns except that the polypyrimidine stretch at the 3` splice junction was less pronounced in the plant introns. A search for possible branch points with sequences related to the yeast, vertebrate and fungal consensus sequences revealed a similar sequence in plant introns.

INTRODUCTION

The interruption of protein coding genes by intervening sequences (IVS, intron) has been observed in all known eukaryoticc genomes. The expression of a large proportion of eukaryotic genes, therefore, requires the excision of introns from messenger RNA precursors (pre-mRNAs) by the process of splicing. The biochemical mechanism of pre-mRNA splicing has been analysed *in vitro* with nuclear extracts from HeLa cells (35-37) and whole cell extracts from the yeast, *Saccaromyces cerevisia* (38), which are able to accurately and efficiently splice exogenously added pre-mRNAs. Pre-mRNA splicing requires the assembly of a ribonucleoprotein complex on the pre-mRNA (spliceosome) (39-43) which is dependent on the U-type small nuclear ribonucleoproteins (snRNPs) (40,41) and on conserved sequences at and near the splice junctions (39-41, 44,45). Following the inital observation that intron sequences started with GT and ended with AG (46) broader splice junction consensus sequences have been derived (47-49).

The elucidation of the biochemical mechanism of splicing al-

so demonstrated that introns are removed as lariat RNAs where the 5` end of the intron forms a 5` - 2` phosphodiester bond with the 2`-OH of an adenosine residue (branch point) lying between 18 and 40 nucleotides from the 3` splice site (50-55). Branch point sequences have been determined for a number of introns allowing the derivation of branch point consensus sequences for yeast, fungal and vertebrate introns. The yeast branch point consensus sequence, TACTAAC, is highly conserved (56) while that of vertebrates and fungia, CTPuAPy (57-59) or PyNPyTPuAPy (51-53) is less highly conserved.

With the exception of the conservation of the GT andAG dinucleotides at the ends of plant introns and the successful splicing in vitro of two plant introns in a HeLa cell nuclear extract (60) little is known about splicing of plant pre-mRNAs. Consensus sequences for plant 3` and 5` splice junctions have been previously derived (17,61). However, these studies were limited by the few plant intron sequences then available (20 introns from 3 gene families of 2 species and 30 introns from 6 gene families of 3 species respectively). In the latter study (61) the introns were analysed for branch point sequences but no consensus similar to that of yeast and vertebrates could be discovered. With the publication in the last two years of genomic sequences of many plant genes, it has been possible to derive splice junction consensus sequences specifically for plant introns (60). In this paper, a catalogue of splice junction and possible branch point sequences is given, the derivation of a plant branch point consensus is presented, and these sequences are compared to those from animal introns.

MATERIALS AND METHODS

The sequences of 167 published and 10 unpublished introns have been collected (1-34) and are presented in Table 1. The plant intron sequences were screened for possible branch point sequences with similar criteria to those used by Keller and Noon (57) in their computer analysis of a variety of animal introns. The region between -15 and -50 from the 3` splice junctions of the plant introns were firstly screened for sequences similar to part of the yeast branch point, CTAAC (56),

Table 1 - Compilation of splice junction and possible branch point sequences from plant introns

Organism and Gene	Line	5` Splice junction	Branch point		3` Splice junction	Ref.
Maize (Zea mays L.)						
Alcohol dehydrogenase, Adh-1	1	AAG: GTCCGC	GCTTGAC	31	CCTGGACCCGTGCAG: C	1
	2	AAG: GTATCT	GGTTGAC	33	CCTTATCTGTCTCAG: G	1
	3	AGG: GTATGT	GCCTGAA	20	TCTTGATTTTGCCAG: T	1
	4	CTG: GTAAGT	TGCTGAG	27	TCTTTCTCTGTTTAG: G	1
	5	GCC: GTAAGT	ATCTGAT	21	CTGCGCATGGTTAAG: G	1
	6	AAG: GTACAG	AGCTCAT	22	TGTCCCATTTTTCAG: C	1
	7	GAG: GTCTGT	TGCTGAA	39	TCCTTTATGGTCTAG: G	1
	8	GAT: GTAAGT	TTCTAAC	21	GCCCTCGTGATCCAG: G	1
	9	AAG: GTAAAT	TGCTGAA	37	TGCAATTCTGCACAG: G	1
Adh-2	10	GAG: GTGCGT	GCCTAAA	38	TGGATCCCTCTGCAG: C	1
	11	AAG: GTCTGT	GCCTAAC	35	TCTTGTCTTGTGCAG: G	1
	12	AGG: GTATGC	AGCTAAC	21	CGCTCTTGGTCGCAG: C	1
	13	CCG: GTAAGC	TACTGAA	25	GTCGTTTTGGTGCAG: G	1
	14	GCT: GTAAGT	CACTCAC	40	TACATGATCGATCAG: G	1
	15	AAG: GTATAA	AACTTAC	26	CTTTTCGTTTTTCAG: C	1
	16	GAG: GTGTGC	ATCTGAT	38	CTGTGTTGCATTGAG: G	1
	17	GAC: GTATGT	GGCTGAA	27	GAAATGGAAATGCAG: G	1
	18	AAG: GTAACC	GACTGAC	45	TGTGTACGTACGTAG: G	1
Glutathione-S-transferase, Gst	19	AAC: GTACCG	CCCTGAC	31	TCTATCTCTCTGCAG: C	2
	20	TCG: GTATGA	TCCTAAT	43	CTGTGTGCTATATAG: A	2
Heat shock protein (70 kD), hsp 70	21	TCG: GTACGC	TACTCAC	30	TTCATTGTAATGCAG: A	3
Sucrose synthetase, shrunken	22	GGG: GTATGC	TGCTGAA	28	TAGCTCGAATTGCAG: T	4
	23	CAG: GTGGGC	ATCTGAG	43	ATACCACTTCTACAG: G	4
	24	CAG: GTAACA	TTCTAAT	21	CTTGTCTGCATATAG: G	4
	25	ACA: GTAAGT	TACTAAT	20	GTCCTTTTTTACCAG: A	4
	26	ACG: GTGAGC	TTCTAAC	20	TGTTTTCTGTTACAG: A	4
	27	TAG: GTGAAT	GATTAAC	32	TATGATCTGTGTTAG: G	4
	28	CAG: GTACAA	TTCTCAT	19	GCAGTCGCTTTGCAG: G	4
	29	ATT: GTATGT	GATTTAC	38	TCTTATTTGTTGCAG: G	4
	30	GAG: GTATAC	TACTGAA	23	CATTCTGTGCTGCAG: G	4
	31	CAG: GTCTGT	GATTAAT	22	TGTACATACTTGCAG: T	4
	32	AAG: GTAGAA	GCTTTAG	48	GTGTTGTTTCTGCAG: C	4
	33	CAA: GTGAGT	AACTGAA	26	TTTACTTGCTTCCAG: G	4
	34	CAG: GTATAT	CACTGAA	37	TTTTTGTGTGGGTAG: C	4
	35	GAA: GTATGC	TCCTGAC	25	CTTTGGATTGCTCAG: G	4
	36	CTG: GTAAGC	TACTGAC	23	CTTTCTGGAATCCAG: G	4
Waxy, wx	37	CAG: GTTCTG	ACCTAAA	41	CTCTCTCCTACGCAG: T	5
	38	GCC: GTAAGC	ATGTGAC	26	CGGGCATGCATGCAG: G	5
	39	GAG: GTACGG	CTCTGAT	26	TGCAAATGCATGCAG: A	5
	40	AGG: GTGAGA	CAGTGAG	36	GGTCGCTGGTTTCAG: G	5
	41	CAG: GTCAGG	CACTGAT	25	CATGCTGTTCTGCAG: G	5
	42	ACG: GTAAGA	CACTGAC	34	CGTCATCCATACAAG: G	5
	43	AAG: GTTGCC	GTCTGAC	21	TTCACGTACTACCAG: A	5
	44	CGG: GTCTGT	ATCTGAC	20	ATTGCATTATTGCAG: C	5
	45	ACG: GTGAGC	TACTGAG	47	TGGTGTCCGGTTCAG: G	5
	46	CTG: GTACGT	GTGTGAG	25	TGGATAATGCTGCAG: G	5
	47	ACG: GTACGA	GATTGAT	29	CTGCGACTCTTGCAG: C	5
	48	GAC: GTAAGC	GTATGAA	45	GTCCTCTCTTCCCAG: T	5
	49	AAG: GTACGT	CGCTGAC	28	TTGCGAAATGCGCAG: G	5
Actin, MAc1	50	AAG: GTTGTT	GCCTAAT	30	CCTCAATATTTACAG: G	6
	51	CTG: GTAAGA	TCCTGAC	34	TATCTCTGTGTGCAG: G	6
	52	CAG: GTCTTC	CACTCAT	47	CAACTGTGTTGGCAG: A	6
Trisephosphate Isomerase	53	TGC: GTAATT	-	-	TCCTGATGCGTGCAG: A	7
	54	TTG: GTACGG	TACTAAA	49	TTGATTGCATTGCAG: A	7
	55	CAG: GTTAGT	AGTTAAT	26	TCATTATTAATGCAG: T	7
	56	GAA: GTATGA	ATCTAAT	29	CTGCTTGGATGGCAG: T	7
	57	CTG: GTACCT	GGCTGAA	29	CTGTTTGTTTTACAG: A	7
	58	GAA: GTAAGT	CGCTCAA	21	GTATTATGTTCCCAG: G	7
	59	GAG: GTACAT	TGCTAAA	40	GCCTCCCTGCTACAG: G	7
	60	AAG: GTAATG	TGCTGAC	28	CTATCTCGTCTGCAG: C	7
Wheat (Triticum aestivum L.)						
Amylase, Amy 13	61	CAG: GTAAGA	GACTGAG	31	TTGTGCGTGCGGCAG: G	8
	62	ATC: GTGAGT	AACTGAT	25	ATTGTGATTCTTCAG: T	8
Amy 18	63	CAG: GTAAGA	TTTTGAT	18	CGAGTTCTGTGGTAG: G	8
	64	ATC: GTGAGT	AACTGAT	25	ATTGTGATTCTTTAG: T	8
Amy 54	65	CAG: GTACGC	TGCTTAA	32	TAATGGATGTTGCAG: G	8
	66	AAG: GTCCCT	CACTAAA	21	TCGACTTGGGTGCAG: G	8
	67	ATC: GTAAGC	TCTTCAA	25	CTCGATGATTTGTAG: T	8
Amy 33	68	CAG: GTGAGA	CTTTCAT	36	TGTTTCGTTGGCCAG: G	8
	69	ATC: GTAAGT	AACTTAC	26	GTTTTGCGCGCGCAG: T	8
Soybean (Glycine max L.)						
Actin, SAc1	70	AAG: GTACAG	CTCTAAC	20	AACGTGTCCTTTCAG: G	6
	71	CTG: GTAAGA	-	-	ATTTTNCTTTTGCAG: G	6
	72	CAG: GTCTGT	TGCTAAT	27	GTCGCTTNAGTGCAG: A	6
SAc3	73	AAG: GTTAGT	AGTTCAT	32	TTTAATATGGAACAG: G	9
	74	CTG: GTTTGT	CCCTGAA	21	TTCCTTTTAAAACAG: G	9
	75	CAG: GTGATT	TGCTAAA	23	GTTGTGGTTTTGCAG: A	9

Table 1(contd.)

Leghaemoglobin, Lb	76	CTC: GTAAGT	TGTTAAT	35	ACTAAAAATGAATAG: G	10
	77	TTG: GTAAGT	TTGTCAC	27	TTTTTTGAATTATAG: G	10
	78	GTG: GTATGA	AGCTAAA	23	CTGATGATTTCGAAG: G	10
Lba	79	TTC: GTAAGT	TGTTAAT	35	ATTAAAAATGAATAG: G	11
	80	TTG: GTAAGT	TCTTCAT	41	TTTTTTGAATTGTAG: G	11
	81	GTG: GTATGA	AGCTAAA	23	CTGATGATTTTGAAG: G	11
Lbc1	82	TTC: GTAAGT	TGTTAAT	35	ATTAAAAATAAATAG: G	11
	83	TTG: GTAAGT	TTGTGAT	23	TTTTCGAATTTGTAG: G	11
	84	GTG: GTATGA	AGCTAAT	31	TTTTATATTTTGTAG: G	11
Lbc2	85	TTC: GTAAGT	ATGTGAG	32	ATTAAAAATTAACAG: G	12
	86	TTG: GTAAGT	TTTTTAT	41	TTTTTTGAATTGTAG: G	12
	87	GTG: GTATGA	AGCTAAT	26	ATGTTTTGTCTGTAG: G	12
Lbc3	88	CTC: GTAAGT	TGTTAAT	35	ACTAAAAATGAATAG: G	12
	89	TTG: GTAAGT	TTGTCAC	27	TTTTTTGAATTATAG: G	12
	90	GTG: GTATGA	AGCTAAA	23	CTGATGATTTCGAAG: G	12
Nodulin, Nod23	91	ATG: GTACGT	TTTTAAT	33	ATTTTGTTGATGCAG: G	13
Nod24	92	AGG: GCAAGT	GGTTCAC	26	GTTAATGTGTTCCAG: C	14
	93	CTG: GTGGTG	ATTTAAT	16	ATTAATGTGTTCCAG: C	14
	94	GTG: GTGGTG	ATTTAAT	16	ATTAATGTGTTCCAG: C	14
	95	GTG: GTGGTG	TACTAAT	17	TTAATGTGTTTGCAG: C	14
Conglycinin, Gmgal7.1	96	GAC: GTAAGC	TCCTTAT	28	CGCTTGATTTTATAG: A	15
	97	GAG: GTAAGT	GATTTAC	25	TGTTCACAAATATAG: G	15
	98	CAG: GTACAT	TTCTAAT	26	ATTGAAAATTTGAAG: G	15
Glycinin, Ala	99	AAG: GTACGT	GATTAAC	35	TGATGTATGGTGCAG: A	16
French bean (Phaseolus vulgaris L.)						
Phaseolin	100	GTG: GTAAGT	TGGTAAT	21	TTTTTATAATTTCAG: G	17
	101	CAT: GTACTG	TTTTAAC	47	ATGTTTGTCCTGTAG: G	17
	102	AAT: GTAAGA	TGTTGAA	37	GCATGATTTTTATAG: A	17
	103	GAG: GTAAAT	ATCTTAG	49	TGTTAACAAATTTAG: G	17
	104	CAG: GTATAT	GCGTGAT	21	ATTGTAAATATGAAG: G	17
Pea, (Pisum sativum L.)						
Legumin, LegA	105	AAG: GTTACT	TACTAAT	27	CTATACCAATTACAG: G	18
	106	AGG: GTGAGC	CAGTAAC	30	ATCTATGTTTGACAG: A	18
	107	AAA: GTATGT	AGCTAAC	22	ACAATCTTCATACAG: A	18
LegD	108	AAG: GTTCGT	TATTTAC	26	TACATCAATTACTAG: G	19
	109	AGG: GTGAGA	-	-	-	19
	110	AAA: GTACCA	GACTTAA	28	ACAATTTTCATACAG: A	19
LegJ	111	AGA: GTAAGT	TACTAAA	30	AATATGTGTATGCAG: G	20
Rubisco, small subunit	112	CAG: GTGACA	TGTTAAC	23	TTGTTGAATATTTAG: G	21
	113	GAG: GTTTCA	CCCTAAT	29	ACTGTTTGGTTGCAG: A	21
Vicia faba L.						
Legumin, LeB4	114	AGA: GTAAGT	AACTCAA	31	ATATGTGTTTTTCAG: G	22
	115	AGG: GTACGT	AACTAAT	35	TGTATGTATATGCAG: A	22
Alfalfa (Medicago sativa L.)						
Glutamine synthetase Gs	116	ATG: GTTAGA	GATTAAT	24	CTCTCATTATGACAG: G	23
	117	AGG: GTAATT	TATTGAT	29	TTTTTTTGGTGCGAG: A	23
	118	CTA: GTATGA	TACTTAT	23	TTGGATTCCTTACAG: C	23
	119	TTG: GTAAGT	GTTTCAT	37	TTTAATTAAATTCAG: G	23
	120	ATG: GTATCT	TTCTGAT	30	ATGATTTGTGATTAG: G	23
	121	CAG: GTGAAA	TTCTAAT	45	TAATTTGCTCAATAG: G	23
	122	CAA: GTAAGT	GTTTAAT	21	GTTTTTTTAATGTAG: T	23
	123	GAG: GTAGGT	AACTAAC	25	TTTATGTTCCAATAG: A	23
	124	AAG: GTTTGC	GTCTTAT	48	TTAATGCAAAACTAG: G	23
	125	CAG: GTAATG	GGTTGAC	26	CTTATAATGCTGTAG: C	23
	126	TGG: GTAAGC	TTCTAAT	29	TTGTGTTATTTGAAG: G	23
Potato (Solanum tuberosum L.)						
Patatin, pgT5	127	CAG: GTATCG	GACTTAT	19	TTCTTTTCGAGTCAG: G	24
	128	TAG: GTACAT	TACTTAT	31	ACATTTATTATGCAG: T	24
	129	AAT: GTAAGT	GACTAAT	26	TTTTTTAAAATGCAG: T	24
	130	CCG: GTACGT	ATCTGAT	34	GTACGTGCAATGCAG: G	24
	131	CAA: GTAAGT	TGCTAAC	25	TATATTTAATTCCAG: G	24
	132	GAG: GTAAAA	TGCTAAC	25	TTTATTTCATTGTAG: G	24
Sb6B	133	TCC: GTAAAA	TTCTGAA	47	TTCTTTTCGAGTCAG: A	24
	134	TGT: GTAGAC	ATTTAAT	27	TATTATATTATGCAG: G	24
	135	AGT: GTAAGT	TTTTAAT	22	TTTAAATGCACGCAG: T	25
	136	TTG: GTAATC	CCCTAAT	31	AACACATGCATGCAG: G	25
	137	CAA: GTAAGT	TGCTAAC	25	TATATTTAATTCCAG: G	25
	138	AAG: GTAAAA	TGCTAAT	25	TTTATTTCGTTGTAG: G	25
SA10C	139	CAG: GTAAAA	GACTCAC	18	TTCTTTTTGCATCAG: G	25
	140	TAG: GTACAT	TACTTAT	33	CATTATATTATGCAG: T	25
	141	TAA: GTCAAA	CACTAAC	28	TAAAAAAAAGTGCAG: T	25
	142	CCG: GTACTA	GTGTGAA	17	TGCTATGCAATGCAG: G	25
	143	CAA: GTAAGT	TGCTAAC	25	TATATTTAATTCCAG: G	25
	144	GAG: GTAAAA	TTCTAAT	25	TTTATTTCGTTGTAG: G	25
pat21	145	CAG: GTATCG	ATCTGAT	49	TTCTTTTCGAGTCAG: G	26
	146	TAG: GTACAT	TACTTAT	31	CATTATCTTATGCAG: T	26
	147	AAT: GTAAGT	GACTAAT	29	TTAAAATGCATGCAG: T	26
	148	CCG: GTACTA	ATCTAAT	26	ACGTACGACGTGCAG: G	26
	149	CAA: GTAAGT	GTCTAAT	21	TATATTTAATTCCAG: G	26
	150	GAG: GTAAAA	TGCTAAT	25	TTTATTTCGTTGTAG: G	26

Table 1(contd.)

Proteinase inhibitor II	151	TTG: GTAAGA	CCTTTAT	19	TATATTTGTTTGTAG: G	27
Carrot (Daucus carrota)						
Extensin	152	AAG: GTACGT	TACTAAA	20	CATATACATTTCGAG: G	28
Tobacco (Niccotiana tabacum L.)						
Rubisco, small subunit	153	CAG: GTAATT	AGCTAAA	25	TTTGGTGGAATATAG: G	29
	154	GAG: GTCAAT	CTTTAAT	22	ATTTTGCATGTGCAG: C	29
	155	CAG: GTCAGT	TCTTGAA	18	CTGGTACTGATGCAG: A	29
Nicotiana plumbaginifolia						
ATP synthase, atp2-1	156	ACC: GTAAGT	GCTTGAT	26	TTCTTGTGGCAACAG: G	30
	157	TTA: GTAAGT	ATCTTAA	21	TTAAAATGGCTACAG: C	30
	158	AAG: GTACTT	TCCTGAT	34	TGTGCTTTTGGTCAG: G	30
	159	ATG: GTTAGG	AGCTGAT	31	GACTATGTTATTCAG: G	30
	160	CAA: GTTAGT	GCCTGAC	26	CCTCAACCATTTCAG: A	30
	161	CAG: GTTGGC	CGCTAAA	27	ATTTTATATTGATAG: G	30
	162	CAG: GTATAA	AACTCAC	45	TCTTTTGGATGCCAG: A	30
	163	CAG: GTAATA	TTTTGAT	29	AATTTCTTTTGACAG: G	30
Antirhinnum majus L						
Chalcone synthase, chs	164	TGT: GTAAGA	TTCTCAC	30	AATTTGAATTATCAG: G	31
	165	CAG: GTACGT	AATTTAT	21	ATTATCCAACACTAG: G	31
Petunia (Mitchell)						
Rubisco, small subunit ssu8	166	CAG: GTACTT	TACTAAT	33	CTCTGTTGAGTATAG: G	32
	167	GAG: GTCAAG	ATCTTAA	23	GTTTTATATGTGCAG: C	32
	168	AAG: GTTAGT	AACTTAG	49	TATGCTCTGTGATAG: G	32
ssu11A	169	CAG: GTACGT	CTTTAGT	39	TTTTGTGGGATGTAG: G	32
	170	GAG: GTTAAG	ATCTTAT	28	GTTTTATATGTGTAG: C	32
Lemna gibba						
Chlorophyl a/b protein	171	CTG: GTTAGA	TGCTCAT	22	GGGCTTCCTGATCAG: G	33
Chlamydomonas reinhardtii						
Rubisco, small subunit, rbcs1	172	CAA: GTTAGT	TTCTAAC	29	ATCGCGTGATCGCAG: G	34
	173	ACG: GTGAGC	ATCTTAC	25	TGCTGTCGCTTGCAG: G	34
	174	TGC: GTAAGT	GACTGAA	36	CCCGTGCGCCCGCAG: C	34
rbcs2	175	CAA: GTGAGT	ATCTAAC	27	CGTTTCCATTTGCAG: G	34
	176	ACG: GTGAGC	CCTTCAT	16	TCCCCTTGCTTGCAG: G	34
	177	TGC: GTAAGT	GACTGAA	36	CCCGTGCGCCCGCAG: C	34

[a]The numbers next to the branch point sequences give the distance in nucleotides of the adenosine branch point nucleotide from the 3` splice junction (:).

and the fungal and vertebrate consensus, CTPuAPy (51,53,58,59). When such sequences were absent the introns were searched for 5 nucleotide sequences with a T in position 2 and an A in position 4. When multiple choices were evident the sequence given in Table 1 was selected by the best fit to the above consensus with the consideration that pyrimidine/purine substitutions represented a bad fit. When more than one sequence of equal fit was present that closest to the 3` splice junction was taken.

RESULTS

Splice junction and possible branch point sequences from forty-three nuclear genes representing twenty-two gene families from fifteen plant species are presented in Table 1. Sequences are presented and discussed in DNA form. The 5` and 3` splice junction sequences are aligned on the basis of the conserved GT and AG dinucleotides, respectively. The frequencies of occu-

Table 2. Nucleotide frequencies at the 5' exon-intron splice junctions of plant introns

Position[a]	-3	-2	-1	:	+1	+2	+3	+4	+5	+6
Total	177	177	177		177	177	177	177	177	177
G	35	19	128		177	0	23	10	115	19
A	58	98	19		0	0	124	98	29	41
C	58	18	19		0	1	13	35	14	30
T	26	42	11		0	176	17	34	19	87
%G	20(9)[b]	11(12)	72(73)		100(100)	0(0)	13(29)	6(12)	65(84)	11(8)
%A	33(40)	55(64)	11(9)		0(0)	0(0)	70(62)	55(68)	16(9)	23(17)
%C	33(43)	10(12)	11(6)		0(0)	1(0)	7(2)	20(9)	8(2)	17(12)
%T	15(7)	24(13)	6(12)		0(0)	99(100)	10(6)	19(12)	11(5)	49(63)
%Pu	53(50)	66(76)	83(82)		100(100)	0(0)	83(91)	61(79)	81(93)	34(25)
%Py	47(50)	34(24)	17(18)		0(0)	100(100)	17(9)	39(21)	19(7)	66(75)
Consensus	C/A	A	G	:	G	T	A	A	G	T

[a] Positions are numbered from the splice site(:). [b] Numbers in brackets are taken from a catalogue of animal intron sequences (49) to allow direct comparison.

rence of the different nucleotides in each position are shown and consensus sequences are derived for the 5` and 3` splice junctions (Tables 2 and 3 and Ref. 60). These values expressed as percentages are also directly compared to those for animal and viral introns (49). The 5` plant splice junction consensus sequence $\begin{matrix}C\\A\end{matrix}$ AG/GTAAGT is virtually identical to that of animal introns $\begin{matrix}C\\A\end{matrix}$ AG/GT$\begin{matrix}A\\G\end{matrix}$AGT. In general, the lower values for the most abundant nucleotides and the higher values of other nucleotides in positions -3,-2, +4, +5 and +6 suggest more variation in the

Table 3. Nucleotide frequencies at the 3' intron-exon splice junctions of plant introns

Position[a]	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	:	+1
Total	176	176	176	176	176	175	176	175	176	176	176	176	176	176	176		176
G	21	25	28	27	25	33	35	41	41	31	19	88	3	0	176		106
A	32	30	23	51	34	37	34	41	36	44	20	35	8	176	0		26
C	40	28	36	20	24	22	33	26	26	23	17	24	118	0	0		24
T	83	93	89	78	93	83	74	67	73	78	120	29	47	0	0		20
%G	12(15)[b]	14(21)	16(10)	15(10)	14(10)	19(6)	20(7)	23(9)	23(7)	18(4)	11(5)	50(24)	2(1)	0(0)	100(100)		60(52)
%A	18(15)	17(10)	13(10)	30(15)	19(6)	21(15)	19(11)	23(19)	20(12)	25(3)	11(10)	20(25)	5(4)	100(100)	0(0)		15(22)
%C	23(19)	16(25)	20(31)	14(21)	14(24)	13(30)	19(33)	15(28)	15(36)	13(36)	10(28)	14(22)	67(65)	0(0)	0(0)		14(18)
%T	47(51)	53(44)	51(50)	44(53)	53(60)	47(49)	42(49)	38(45)	41(45)	44(57)	68(58)	16(29)	27(31)	0(0)	0)0)		11(8)
%Pu	30(30)	31(31)	29(29)	44(26)	34(16)	40(21)	39(18)	47(28)	44(19)	43(7)	22(15)	70(49)	6(4)	100(100)	100(100)		75(74)
%Py	70(70)	69(69)	71(71)	56(74)	66(84)	60(79)	61(82)	53(72)	56(81)	57(93)	78(85)	30(51)	94(96)	0(0)	0(0)		25(26)
Consensus	T	T	T	T[c]/Pu	T	T	T/Pu	T/Pu	T/Pu	T/Pu	T	G	C	A	G	:	G

[a] Positions are numbered from the splice site(:). [b] Numbers in brackets are taken from a catalogue of animal intron sequences (49) to allow direct comparison. [c] At these positions T is the most abundant single nucleotide but the combined %G and %A are greater than or very similar to the %T.

Table 4. Comparison of the pyrimidine/purine content of the polypyrimidine stretch at the 3' splice site between animal and plant introns.

	Animal/Viral[a]	Plant
Total number of introns examined	124	176
Introns with 5 or more consecutive pyrimidines in positions -5 to -15	80(65%)	36(20%)
Introns with 7 or more consecutive pyrimidines in positions -5 to -15	51(41%)	15(9%)
Introns with 0,1 or 2 purines in positions -5 to -15	80(65%)	22(13%)
Introns with 5 or more purines in positions -5 to -15	9(7%)	54(31%)

[a] Values are derived from Mount(1982) but do not include the plant introns presented in that study (49).

plant intron sequences. At position +3 in the plant consensus sequence the occurrence of G residues is lower and that of A residues is slightly higher.

The plant 3` consensus sequence, $TTT^{T}_{Pu}TT^{T\ T\ T\ T}_{PuPuPuPu}TGCAG/G$, differs from that of animals in that, firstly, at position -4 a G occurs while any nucleotide (N) can occur in the animal sequence, and secondly, the polypyrimidine stretch at positions -5 to -15 is much less pronounced (Table 3). The occurrence of purines is increased in the plant sequences such that the range of percentage purines increases in plants to 22 to 47% as compared to animal and viral sequences, 7 to 31% (49). Although in all positions (-5 to -15) thymidines are the most abundant, the percentage purines is greater than or equal to the % thymidine in positions -7, -8, and -12 and only slightly less than the percentage thymidines in positions -6 and -9. In virtually all positions the % cytidine is greatly reduced when compared to the animal intron values. The higher occurence of purines in positions -5 to -15 is most clearly seen when the plant intron sequences in Table 1 and the animal and viral intron sequences (49) were analysed for the number of purines and for the occurrence of stretches of consecutive pyrimidines (Table 4). Only 20% of the plant introns contained a stretch of 5 or more consecutive pyrimidines in positions -5 to -15 and only 9% contained 7 or more consecutive pyrimidines. On the other hand 65% and 41% of the animal and viral sequences (49) contained 5 or more and 7 or more consecutive pyrimidines respectively, in these positions (Table 4).

Twenty-three percent of the animal sequences contained 9, 10

Table 5. Nucleotide frequencies at putative branch points in plant introns

Position[a]	-5	-4	-3	-2	-1	0	+1
Total	174	174	174	174	174	174	174
G	34	43	11	0	61	0	9
A	45	51	1	0	68	174	41
C	25	22	120	0	22	0	48
T	70	58	42	174	23	0	76
%G	20	25	6	0	35	0	5
%A	26	29	1	0	39	100	24
%C	14	13	69	0	13	0	28
%T	40	33	24	100	13	0	44
%Pu	45	54	7	0	74	100	29
%Py	55	46	93	100	26	0	71
Consensus	T[b] Pu	T Pu	C[c] T	T	Pu	A	Py

[a]Positions are numbered from the branch point nucleotide (0). [b]See Table 3.
[c]At this position there is a much higher frequency of C's than T's.

or 11 consecutive pyrimidines while only three plant introns (2%) contain 9 consecutive pyrimidines and none contained 10 or 11. Nine of the animal intron sequences (7%) contained 5 or more purines in positions -5 to -15 of which only one intron contained as many as 7 purines. On the other hand thirty-one percent of the plant introns contained 5 or more purines of which two contained 7 purines, four contained 8 purines, and two contained 9 purines in the eleven positions (-5 to -15). The frequencies of occurrence of nucleotides of the possible branch point sequences is shown in Table 5 and a consensus sequence is derived: CTPuAPy. This sequence is identical to the fungal and vertebrate branch point consensus sequence (51,53,57,58). A number of the plant introns contained more than one potential branch point sequence and that given in Table 1 represents the best fit to the criteria given in the Materials and Methods section.

DISCUSSION

The plant 5` splice junction consensus sequence (Table 2) is virtually identical to that of animals. Of the 177 intron sequences present only the first intron of the nodulin-24 gene from soybean does not confer to the GT rule but instead starts with GC (14). Besides this violation of the GT rule in the

first intron, the nodulin -24 gene has an unusual gene structure in that the second, third and fourth introns are virtually identical having been formed by the direct repetition of a 200 bp intron containing sequence. Although this feature is apparently not an artefact and the gene is apparently expressed this single violation of the GT rule requires further investigation.

The plant 3` splice junction consensus sequence (Table 3) (Table 3) is similar to that of animals, $(^{T}_{C})_{11}$NCAG/G (49) with two exceptions. Firstly, at positions -4 the plant sequence has a G instead of any nucleotide (N). Secondly the polypyrimidine stretch at positions -5 to -15 is not as pronounced in the plant sequences. The polypyrimidine stretch has been shown to be necessary for spliceosome assembly and, therefore, for splicing in the HeLa cell *in vitro* splicing system (39,53). However, the exact requirement in terms of number and positioning of pyrimidines is still unknown. This difference between the plant and animal 3` splice junctions may reflect a difference in one or more of the factors required for mRNA splicing.

The concensus of possible branch point sequences from plant introns is identical to that of animals, CTPuAPy. However, since the nature of plant branch points is unknown, none having been determined in homologous *in vitro* or *in vivo* systems, this consensus must be taken tentatively. Branch point sequences from introns of an amylase gene of wheat and a legumin J gene of pea have been mapped in the HeLa cell *in vitro* splicing system and the sequences show a good fit to the branch point consensus (60). None of the introns in Table 1 contain the highly conserved TACTAAC sequence of yeast.

ACKNOWLEDGEMENTS

This work was supported by grants to G. Feix of this department from the Deutsche Forschungsgemeinschaft and the Fonds der Chemischen Industrie.

REFERENCES

1. Dennis, E.S., Sachs, M.M., Gerlach, W.L., Finnegan, E.J. and Peacock, W.J. (1985) Nucl. Acids Res. 13, 727-743.
2. Shah, D.M., Hironaka, C.M. Wiegand, R.C., Harding, E.I., Krivi, G.G. and Tiemeier, D.C. (1986) Plant Mol. Biol. 6, 203-211.

3. Rochester, D.E., Winer, J. A. and Shah, D.M. (1986) EMBO J. 5, 451-458.
4. Werr, W., Frommer, W.-B., Maas, C. and Starlinger, P. (1985) EMBO J. 4, 1373-1380.
5. Klösgen, W.B., Gierl, A., Schwarz-Sommer, S. and Saedler, H. (1986) Mol. Gen. Genet. 203, 237-244.
6. Shah, D.M., Hightower, R. C. and Meagher, R.B. (1983) J. Mol. Appl. Genet. 2, 111-126.
7. Marchionni, M. and Gilbert, W. (1986) Cell 46, 133-141.
8. D. Baulecombe, in preparation.
9. Shah, D. M. Hightower, R.C. and Meagher, R.B. (1982) Proc. Natl. Acad. Sci. USA 79, 1022-1026.
10. Brisson, N. and Verma, D.P.S. (1982) Proc. Natl. Acad. Sci. USA 79, 4055-4059.
11. Hyldig-Nielsen, J.J., Jensen, E.O., Paludan, K., Wiborg, O., Garrett, R., Jorgensen, P. and Marcker, K.A. (1982) Nucl. Acids Res. 10, 689-701.
12. Wiborg, O., Hyldig-Nielsen, J.J., Jensen, E.O., Paludan, K. and Marcker, K.A. (1982) Nucl. Acids Res. 10, 3487-3493.
13. Mauro, V.P., Nguyen, T., Katinakis, P. and Verma, D.P.S. (1985) Nucl. Acids Res. 13, 339-349.
14. Katinakis, P. and Verma, D.P.S. (1985) Proc. Natl. Acad. Sci. USA 82, 4157-4161.
15. Schuler, M.A., Schmitt, E. and Beachy, R.N. (1982) Nucl. Acids Res., 10, 8225-8244.
16. Marco, Y.A., Thanh, V.H., Tumer, N.E., Scallon, B.J. and Nielsen, N.C. (1984) J. Biol. Chem. 259, 13436-13441.
17. Slightom, J.L., Sun, S.M. and Hall, T.C. (1983) Proc. Natl. Acad. Sci. USA 80, 1897-1901.
18. Lycett, G.W., Croy, R.R.D., Shirsat, A.H. and Boulter, D. (1984) Nucleic Acids Res. 12, 4493-4506.
19. Bown, D., Levasseur, M., Croy, R.R.D., Boulter, D. and Gatehouse, J. A. (1985) Nucl. Acids Res. 13, 4527-4537.
20. Gatehouse, J.., in preparation.
21. Coruzzi, G., Broglie, R., Edwards, C. and Chua, N.-H. (198o4) EMBO J. 3, 1671-1679.
22. Bäumlein, H., Wobus, U., Pustell, J. and Kafatos, F.C. (1986) Nucl. Acids Res. 14, 2707-2720.
23. Tischer, E, DasSarma, S. and Goodman, H.M. (1986), Mol. Gen. Genet. 203, 221-229.
24. Rosahl, S., Schmidt, R., Schell, J. and Willmitzer, L. (1986) Mol. Gen. Genet. 203, 214-220.
25. Pikaard, C.S., Mignery, G.A., Ma, D.P., Stark, V.J. and Park, W.D. (1986) Nucl. Acids Res. 14, 5564-5566.
26. Bevan, M. Barker, R., Goldsbrough, A., Jarvis, M., Kavanagh, T. and Iturriaga, G. (1986) Nucl. Acids Res. 14, 4625-4638.
27. Keil, M., Sanchez-Serrano, J., Schell, J. and Willmitzer, L. (1986) Nucleic Acids Res. 14, 5641-5650.
28. Chen, J. and Varner, J.E. (1985) EMBO J. 4, 2145-2150.
29. Mazur, B.J. and Chui, C.-F. (1985) Nucl. Acids Res. 13, 2373-2386.
30. Boutry, M. and Chua, N.-H. (1985) EMBO J. 4, 2159-2165.
31. Sommer, H. and Saedler, M. (1986) Mol. Gen. Genet. 202, 429-434.
32. Tumer, N.E., Clark, W.G., Tabor, G.J., Hironaka, C.M.,

Fraley, R.T. and Shah, D.M. (1986) Nucl. Acids Res. 14, 3325-3342.
33. Karlin-Neumann, G.A., Kohorn, B.D., Thornber, J. P. and Tobin, E.M. (1985) J. Mol. Appl. Genet. 3, 45-61.
34. Goldschmidt-Clermont, M. and Rahire, M. (1986) J. Mol. Biol. (in press).
35. Hernandez, N. and Keller, W. (1983) Cell 35, 89-99.
36. Hardy, S.F., Grabowski, P.J., Padgett, R.A. and Sharp, P.A. (1984) Nature 308, 375-377.
37. Krainer, A.R., Maniatis, T., Ruskin, B. and Green, M.R. (1984) Cell 36, 993-1005.
38. Lin, R.J., Newman, A.J., Cheng, S.-C. and Abelson, J. (1985) J. Biol. Chem. 260, 14780-14792.
39. Brody, E. and Abelson, J. (1985) Science, 228, 963-967.
40. Frendewey, D. and Keller, W. (1985) Cell 42, 355-367.
41. Grabowski, P.J., Seiler, S.R. and Sharp, P.A. (1985) Cell 42, 345-353.
42. Bindereif, A. and Green, M.R. (1986) Mol. Cell Biol. 6, 2582-2593.
43. Kaltwasser, G., Spitzer, S.G. and Goldenberg, C.J. (1986) Nucl. Acids Res. 14, 3687-3701.
44. Ruskin, B. and Green, M.R. (1985) Cell 43, 131-142.
45. Vijayraghavan, U., Parker, R., Tamm, J., Iimura, Y., Rossi, J., Abelson, J. and Guthrie, C. (1986) EMBO J. 5, 1683-1695.
46. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349- 383.
47. Rogers, J. and Wall, R. (1980) Proc. Natl. Acad. Sci. USA 77, 1877-1879.
48. Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.M. and Steitz, J.A. (1980) Nature 283, 220-224.
49. Mount, S.M. (1982) Nucleic Acids Res. 10, 459-472.
50. Padgett,R.A., Konarska, M.M., Grabowski, P.J., Hardy, S.F. and Sharp, P.A. (1984) Science 225, 898-903.
51. Ruskin, B., Krainer, A.R., Maniatis, T. and Green, M.R. (1984) Cell 38, 317-331.
52. Konarska, M.M., Grabowski, P.J., Padgett, R.A. and Sharp, P.A. (1985) Nature 313, 552-557.
53. Zeitlin, S. and Efstratiadis, A. (1984) Cell 39, 589-602.
54. Reed, R. and Maniatis, T. (1985), Cell 41, 95-105.
55. Ruskin, B., Greene, J.M. and Green, M.R. (1985) Cell 41, 833-844.
56. Teem, J., Aborisch, N. Kaufer, N. Schwindinger, W., Warner, J., Levy, A., Woolford, J., Leer, R., Van Raamsdonk-Duin, M., Mager, W., Planta, R., Schultz, L., Friesen, J., Fried, H. and Robash, M. (1984) Nucl. Acids Res. 12, 8295-8312.
57. Keller, E.B. and Noon, W.A. (1984) Proc. Natl. Acid. Sci. USA 81, 7417-7420.
58. Kinnaird, J.H. and Fincham, J.R.S. (1983) Gene 26, 253-260.
59. Käufer, N.F., Simianis, V., and Nurse, P. (1985) Nature 318, 78-80.
60. Brown, J.W.S., Feix, G. and Frendewey, D. (1986) EMBO J. (in press)
61. Rogers, J. M. (1985) Int. Rev. Cytol. 93, 188-279.