

MINIREVIEW

Regulatory sequences in the transcription of *Neurospora crassa* genes: CAAT box, TATA box, Introns, Poly(A) tail formation sequences

Jon J.P. Bruchez, J. Eberle and V.E.A. Russo - Max Planck Institut für Molekulare Genetik, Ihnestr. 73, D-14195 Berlin, Germany

We have analyzed the sequences of 77 nuclear genes of *N. crassa* thought to be transcribed by RNA polymerase II (references 1-72 of the accompanying paper). In Table I we present the data on regulatory sequences in the 5' region, in Table II the data on the regulatory sequences located at the 3' end of the genes, in Table III the regulatory sequences involved in intron splicing of *N. crassa* genes, and, in Table IV, a study of the distribution of these introns.

While in mammalian systems the binding proteins for at least two regulatory sequences, the CAAT and TATA boxes (Montague, 1987, *Gene Structure in Eukaryotic Microbes*, Kinghorn, Ed., IRL Press p. 263), are known, in *N. crassa* no binding proteins for any such sequences have yet been identified. The validity of the boxes presented here is therefore based entirely upon statistical analysis. Note, though, that Selker *et al.* (1986) *Mol. Gen. Genet.* **205**:189-192) have shown through deletion analysis that a (A/T)TATA(A/G) box, highly conserved in both sequence and position, appears to play a role in the regulation of transcription of the 5S rRNA genes of *N. crassa*. This is unusual as such a sequence is not usually associated with other Pol III transcribed genes.

When is a box statistically significant?

N. crassa has a G+C content of 54% (Villa and Storck, 1968, *J. Bacteriol.* **96**:184-190); we assume 50% here for simplicity in our calculations. Whether a box is statistically significant or not depends both on the length of the sequence, on its stringency to a defined consensus and the window (expressed in bp) in which it is to be found. A box of 4 given bases, no matter whether contiguous or not, has a probability of 1 in 256 to be found in a region of DNA just large enough to house the particular sequence, a window of 1 bp. We define a window as the number of possible positions that each base in the box is permitted to occupy on the DNA sequence. It will be found with a probability of $256/256 = 100\%$ in a window of 256 bp. Given below are the probabilities to find boxes of given length and stringency in a 1 bp window:

$4/4 = 1/256$	$5/5 = 1/1024$	$6/6 = 1/4096$	$7/7 = 1/16384$
$4/5 = 1/68$	$5/6 = 1/227$	$6/7 = 1/781$	
	$5/7 = 1/86$		

Regulatory sequences in the 5' region**The CAAT box**

Fifty genes had determined 5' mRNA ends (+1) and were screened for possible CAAT type boxes around the -80 bp position, the usual location for mammalian CAAT boxes. When several 5' ends were given, +1 was taken to be the most distal from the ATG except when the authors indicated the major site themselves. Six genes were found to harbor a CAAAT sequence (underlined in Table I) in a range of -75 to -88, a window of 13 bp. The cumulative window of 13 bp in 50 genes is 650 bp. The probability of finding a 5/5 box in a 1 bp window is $1/1024$, therefore in a window of 650 bp, the box should occur $650/1024 = 0.6$ times. In other words, on a statistical basis we should find the CAAAT sequence in this position only 0.6 times

out of all 50 genes. The fact that we find 6 indicates that the CAAAT box is of statistical significance, 10 times above statistical background. For comparison the mammalian CAAT box has a consensus of GG(C/T)CAATCT at around -80 bp. (Montague, 1987, Gene Structure in Eucaryotic Microbes, J. Kinghorn Ed., IRL Press, P. 263)

The TATA box

The genes with a defined +1 were also screened for the presence of a possible TATA box around 30 bp upstream of the +1. This revealed a statistically highly significant TATATAA box which is present in 5 of the genes (double underlined in Table I) at a distance from the first +1 of 34-44 bp (window 10 bp). The probability that 1 gene out of 50 has such a sequence is $(10 \times 50)/16384 = 0.03$. We have found 5 such genes giving a factor of 150 above statistical background. If there is indeed a factor that can bind to this sequence even with one wrong base then there are 6 more genes with a degenerate TATATAA (single underlined in Table I). For comparison, the mammalian and yeast TATA box is: TAT(A/T)A(A/T) at around -30 bp (Montague, 1987, Gene Structure in Eucaryotic Microbes, J. Kinghorn Ed., IRL Press, P. 263)

+1 Sequence Consensus

The most striking consensus sequence around the +1 is TCATCANC (double underlined in Table I) which has a probability of $1/16384$ of being found in a 1 bp window as a 7/7 sequence. There are 6 genes in which 16 transcription starts lie either within the TCATCANC sequence itself or up to two bases away so that we can consider the window to be 12 bp. The probability that out of 50 genes, there is one gene with at least one +1 (there are 107 +1's highlighted in Table I) lying within a 12 bp window of such a sequence is $(107 \times 12)/16384 = 0.08$. Having found 6 genes with 16 transcription starts, that gives us a factor of 200 over background. Similar results are obtained if we consider only the first transcription start point. We have scored 5 other genes with this sequence at single base degeneration (single underlined in Table I).

Regulatory Sequences in the 3' Region

Polyadenylation signal sequences

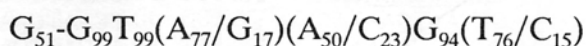
Among the 77 genes in consideration there are 29 which have a 3' end of their mRNA determined by the poly(A) of their cDNA. Some have several 3' ends, so in total there are 34 3' ends given. The polyadenylation sequence in mammals is AATAAA, and in yeast is AATAA, both located 10 to 30 bp upstream of the poly(A) tail (Montague, 1987, Gene Structure in Eucaryotic Microbes, J. Kinghorn Ed., IRL Press, P. 263). We looked within the same region and found two genes, *nit-4* and *spe-1*, with the AATAAA sequence 20 and 17 bases respectively upstream from their 3' ends (double underlined in Table II). Statistically we expect $(34 \times 20)/4096 = 0.16$ such sequences in a window of 20 bases. Therefore 2 is 12 times more than expected. There are then 14 more 3' ends showing the same sequence 3 to 23 bp upstream, window 20 bp, but with a stringency of 5/6 (single underlined in Table II). Statistically we expect $(34 \times 20)/227 = 4$ so there is a factor of 3 over statistical background.

Intron Regulatory Sequences

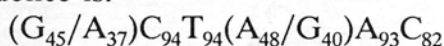
Table III presents the introns from the 77 genes analyzed. There are in total 149 introns with a distribution similar to the Poisson distribution (Table IV). The number of introns seems to be quite independent of the length of the gene, for example, the three genes with 7 introns,

atp-2, *crp-1*, and *nur-40* have coding regions of 2200 bp, 950 bp, and 1600 bp, respectively, including the introns. Among the genes without introns there are some very large, e.g. *frq* 2360 bp, *nuc-1* 2565 bp, *qa-1F* 2400 bp, and some very short or middle length, e.g. *cys-3* 710 bp, *met-7* 1630 bp, *qa-4* 1100 bp. Of the 14 genes without introns there are 5 in the *qa* cluster of 7 genes. This indicates that the location of the gene might be important in determining whether or not it contains introns.

The 5' signal is:

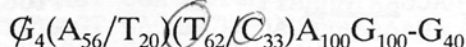


The Lariat or internal sequence is:



with a distance of 6 to 29 bases from the 3' splice site.

The 3' signal is:



where the subscript number indicates the % occurrence of the particular nucleotide, $\emptyset$ indicates a conserved absence of that particular nucleotide, and - shows the splicing site.

We have determined the majority of the internal lariat sequences presented here.

These *N. crassa* intron signals are very similar to the mammalian signals which are:

5': AG-GT(A)AGT
Lariat: CT(A/C)A(T/C)?
3': (T/C)₁₁NCAG-G

(Montague 1987 Gene Structure in Eucaryotic Microbes, J. Kinghorn Ed., IRL Press, P. 263)
For *N. crassa*, intron lengths lie between 46 and 856 with a tendency toward 60 bp.

Table I Regulatory sequences in the 5' region of *Neurospora crassa* genes

Gene	CAAT box	Dist C ^a	TATA box	Dist T ^b	TCATCANC at +1
<i>am</i>	-	-	TCTTGTATAAAGT	38	-
<i>arg-2</i>	AATCAAATGTC	85	-	-	CCGTCATCAACTCT
<i>atp-1</i>	-	-	-	-	CCTCCATCAACCTCCCGCTACATCT
<i>atp-2</i>	GAGCAAATCAC	78	-	-	-
<i>bli-7</i>	-	-	GTGTATATAAGAC	42	TCATCATCAGCATC
<i>chs-1</i>	-	-	-	-	CTAGCATCATCTAG
<i>cmt</i>	-	-	GGGTATATAAAGC	44	TTGTCAATCAACCGA
<i>con-10</i>	-	-	GTGTATATAAGCA	42	-
<i>con-13</i>	-	-	ATGCATATAAGAA	30	-
<i>cys-3</i>	-	-	TTGTATATCAGAT	37	-
<i>for</i>	-	-	-	-	CCTTCATCATCCTC
<i>grg-1</i>	-	-	GCCTATATAAGAC	42	CCATCATCAGCCAA
<i>his-3</i>	TACCAAATCAC	88	CTGTACATAAGCG	46	-
<i>hsp30</i>	-	-	TCAAATATAAATC	46	-
<i>laccase</i>	-	-	ACGTGTATAAAGT	45	TCTTCATCATCATA
<i>lox</i>	-	-	GTCTATATAAGAG	34	-
<i>pho-4</i>	-	-	-	-	TTCTCTTCAGCACC
<i>qa-4</i>	GGTCAAATCAAATCTT	88	-	-	ATTTCCTCACCATT
<i>qa-x</i>	CAGCAAATGCT	75	-	-	-
<i>spe-1</i>	TCACAAATTTTC	81	-	-	-
<i>T</i>	-	-	-	-	ACTACATCAGCAGT

Key: Double Underlining indicates a box showing perfect stringency.

Single Underlining indicates a box showing a single nucleotide degeneracy.

a is the distance from the end of the CAAT box to the first major +1

b is the distance from the end of the TATA box to the first major +1

Nucleotides shown in *italics* are major +1 transcription initiation sites.

Table II Regulatory sequences located at the 3' end of *Neurospora crassa* genes among the 29 genes with 3' determined by poly(A) tail in their cDNA:

Gene	Dist. 3 ^a	AATAAA	Dist. A ^b	3' terminal sequence
<i>acp</i>	218	TGTA <u>AATACA</u> AGA	19	GCTGTTTCCCCATGTGTATTC
<i>al-1</i>	128	GGGTATAAA <u>CGA</u>	14	TTCGTAGATAAAGTCTTGGGA
	160	GGGAATATATAG	15	GTTTGTGTTTATGTGGAAGA
<i>atp-1</i>	27	TTGAATTAATTC	7	CTATTCCCGTGGTTCTTGAGA
<i>atp-2</i>	73	TAGAGTAAAGAA	15	GGTTTTTCGGGACGTTCTCTCC
	194	GTTAATGAATAC	17	TGAATGAACCTCAGCCTATGTG
<i>cys-3</i>	540	TGAAAGAAAAAGA	23	ACGAGTCATGCAAAAAAGAAG
<i>grg-1</i>	223	GCCAAATTAATAC	3	TTAATACCTTCACATCTGTTT
<i>hsp30</i>	374	ATAACTAAAGTT	22	GAACACAACCTACGCCAGTTCC
<i>ilv-2</i>	327	GTCAATGAACTT	21	TTTTCTCTGTCAATGGCTTG
<i>nit-3</i>	50	AGCAATGAATTG	17	CCTTCAGATGACCTTTTGTGT
<i>nit-4</i>	129	CACAATAAAATGC	20	GTCGTTCTATACACAAATTC
<i>nur22</i>	266	TCCAAATAACATT	15	TCTCCTTCTTGAATCATCAT
<i>nur40</i>	135	AGCAATCAAGAG	9	AGAGATTGCGCAACGTTTGA
<i>spe-1</i>	201	ACGAATAAAATT	17	TTGGACCCTATAAGATATTTG
T	430	ACAAAAAAGAC	18	TTTTTCATCGCCGTAACCACC

Key: Double Underlining indicates a AATAAA box showing perfect stringency.

Single Underlining indicates a AATAAA box showing a single nucleotide degeneracy or poly(A) tail addition sites. .

a is the distance between last codon and first poly(A) site.

b is the distance between AATAAA consensus and first poly(A) tail site.

al-1 and *atp-2* each have two poly(A) tail addition sites.

Table III Regulatory sequences involved in intron splicing in *Neurospora crassa* genes.

5' Consensus: G₅₁-G₉₉T₉₉(A₇₇/G₁₇)(A₅₀/C₂₃)G₉₄(T₇₆/C₁₅)
 Lariat Consensus: (G₄₅/A₃₇)C₉₄T₉₄(A₄₈/G₄₀)A₉₃C₈₂
 3' Consensus: G₄(A₅₆/T₂₀)(T₆₂/C₃₃)A₁₀₀G₁₀₀/G₄₀

Ref.	Gene	Id ^a	5' signal	Lariat signal	Dist ^b	3'signal	L ^c
			G GTRNGY	RCTRAC	6-29	N YAG G	
1	<i>acp</i>	D	1) CCGOGTATGT 2) CAGOGTAGGT	ACATGCTAACATCGC GTGTGCTGACGACCC	18 10	CTACAG CCC CCCTAG GAT	380 192
2	<i>acu-3</i>	D	1) GCTOGTTAGT or ACAATCTCACTCGGC 2) TACOGTGAGT or AGCCCCTCCCATACT or CCATACTGATATTCC	ACAATCTCACTGACA ACAATCTCACTCGGC AGCCCCTCCCATACT CCATACTGATATTCC	21 10 18 x 12 x	CGACAG AAG ATCTAG ACC	70 66
3	<i>acu-5</i>	S	1) ACT1GTAAGT	AGATACTAACAGCTG	12	AAATAG CTC	58
4	<i>acu-8</i>	D	1) CAA1GTAAGT 2) TACOGTAAGT	AGTTGCTAACCCATG GTTTGCTAACCCCTA	13 20	CTACAG GAA CAACAG GGC	73 67
5	<i>acu-9</i>	S	1) CAGOGTATGT	TCATACTAACAACCA	11	CAACAG GAG	46
6	<i>al-1</i>	D	1) TTG1GTATGT 2) TTCOGTAAGT	GCTAACTTCTTCCCC TCCAATAACTTCAC	15 x 21	CAACAG GCG GAACAG TAC	77 108
7	<i>al-3</i>		NO INTRONS				
8	<i>alc</i>	D	1) TCG1GTCCGT	TTCAACTAACGGAAG	21	ATACAG ATC	72
9	<i>am</i>	C	1) AGG1GTACGT 2) GCC1GTAAGT	CAGAGCTGACTTGAT ATTTGCTGACTCGGC	17 13	CCACAG AGT CTCTAG TGA	67 61

Ref.	Gene	Id ^a	5' signal	Lariat signal	Dist ^b	3'signal	L ^c
10	<i>arg-2</i>	D	1)	AGGOGTGCGT TAAACCTAACATTTT	14	GCTCAG GAT	56
11	<i>atp-1</i>	D	1)	GCGOGTAGGT TAGGGCTAACTCGAC	8	CAGCAG CGA	202
			2)	CGG1GTACGT GGATGCTGACGTGTC	16	GTATAG TGA	309
			3)	CGA2GTATGT CAAATCTGACCCTTT	13	CCCCAG GTT	63
			4)	GGTOGTAAGT ACTGGCTAACCAGAA	18	ACACAG GCG	323
			5)	CGTOGTAAGT CAGAGCTGACGAGTC	14	CTACAG TTG	61
11	<i>atp-2</i>	D	1)	GAG2GTGAGT TTGGCCTTCCTCTTG	16 x	ATATAG CGG	111
			2)	TTG1GTAAGC CCTTGCTAACCGCGC	21	CCACAG GTG	157
			3)	ATG1GTCAGT TTATACTGACCCCGC	18	CAACAG TCA	101
			4)	TCGOGTACGT ATGCGCTAACCAGCC	11	CCGCAG CAA	88
			5)	ACA1GTAAGC ATTCGCTGACATGAT	17	TTATAG CTG	69
			6)	CCC2GTGGGT TTTTACTGACGCAAA	12	GTGTAG TGT	83
			7)	ATG1GTATGT CGTTGCTAACGCAGT	11	CTGTAG TGT	61
12	<i>bli-7</i>	D	1)	AACOGTAAGT CCTTGCTAACCTTCG	26	AAAAAG ACC	95
13	<i>Bml</i>	D	1)	ATTOGTAAGT CGACGCTGACACGAT	21	CTATAG GTT	240
			2)	TGCOGTAAGT CAGGACTAACACAAC	17	GATCAG GGT	74
			3)	CTG2GTACGT CGACGCTGACAGAAT	11	AAACAG GCA	68
			4)	TGT2GTACGT GAAAGCTCACC GCCC	12	CTACAG GTA	66
			5)	GAGOGTGAGC GCTCGCTAACTAGCT	18	TGACAG GCT	73
			6)	CTT2GTAAGT TAATACTGACGAATC	11	AAACAG CCG	57
14	<i>chs-1</i>	D	1)	CAG1GTAAGT TAACACGAACGTCGT	12 x	ATCCAG GGG	73
			2)	GGG2GTAAGC AACCCTTACTAATA	16	TGATAG CAA	59
15	<i>cmt</i>	D	1)	GCT1GTAAGT TGGTACTAACTTTGA	15	TTCTAG GCT	94
16	<i>con-8</i>	D	1)	CGGOGTATGT ATGTGCTAACAGCTC	23	ACATAG CCA	169
			2)	TAA2GTACGT TTAAGCTAACTCGTT	17	TAATAG TTG	69
17	<i>con-10</i>	D	1)	CCG2GTATGT CTTTGCTAACATAAT	17 x	CTCCAG CCC	70
			2)	CAGOGTATGT GTTGACCAACACATG	17	AAACAG CGC	74
18	<i>con-13</i>	D	1)	GATOGTAGGT CTGTGCTTACCCTTAA	16	CAATAG TGC	57
			2)	GGA2GTAAGT CTGTGCTGACCGGAA	14	AAACAG CAC	62
19	<i>cot-1</i>	C	1)	CCA2GTATGC TCATTCTAACATTGA	14	TACTAG CAA	78
			2)	CAG2GTAAGC AGATACTGACACGGT	16	ATGCAG AGA	59
			3)	AAG2GTATGC ACGCGCTCACCATAT	18	TCATAG CCT	58
20	<i>cpc-1</i>	D	1)	CAG1GTAATT ATGCGCTTACAATCT	12	GCACAG AAC	57
21	<i>cpi</i>	S	1)	CCG?GTACGT ATTGGCTGACCCCTC	18	TTTTAG TGA	856
			2)	GGT?GTAAGT ACCGACTGACCTGCA	13	CTTTAG TTT	94
			3)	AGG?GTAAGT GATGTCTAACTCCCA	11	ATGCAG CTC	271
			4)	AAC?GTAAGT AAGACCTAACCTCTC	12	GAACAG GGG	66
22	<i>crp-1</i>	D	1)	ATGOGTATGG AAACGCTGATTCACT	15 x	ATGTAG CCT	47
			2)	CATOGTAAGC GATGACTGACTGTAG	16	TTATAG GTT	50
			3)	GGG1GTATGT GAGTATTGACAGCAT	13 x	TTCCAG CCG	62
			4)	AGG2GTACGT TGATGCTAACAATGG	11	GAACAG TGG	62
			5)	CAGOGTATGC CGATACTAACCCGAC	11	GATAAG CAC	63
			6)	AGT1GTACGC AAACGCTGACGATGA	12	GGATAG ACC	126
			7)	AGG1GTAAGA CTCGTCTAACAACAC	12	TTCTAG GCC	61
23	<i>crp-2</i>	C	1)	GCG1GTAAGT GGAGGCTGACAATCA	11	ATTTAG TTG	73
			2)	CAG2GTTTCGT TGAGGCTAACATCCT	17	TTCCAG TGG	215
			3)	GTG1GTATGT TGATCCTAACATTTT	10	TCATAG TCA	54
24	<i>crp-3</i>	D	1)	AAGOGTGCGT GGGATCTAACATGTT	17	CAATAG ATT	93
			2)	GGCOGTAAGT TTTGTCTAACTTACC	14	GAACAG TTC	98
			or	TCTAACTTACCTTCG	10		
25	<i>cya-4</i>	D	1)	CTG1GTAAGT AAAGACTGACATGTA	21	ACGCAG CCT	398
			2)	AAGOGTGCGT ATCAACTAACACATA	21	AAACAG GCC	68
26	<i>cys-3</i>	NO	INTRONS				

Ref.	Gene	Id ^a	5' signal	Lariat signal	Dist ^b	3' signal	L ^c
27	<i>cys-14</i>	D	1) TCC2GTTTGT 2) AAT2GTATGG 3) GTG1GTAAGT 4) GGTGTTACGT	AGATACTGACAAGAT TATTGCTAACATAAT CGTAACTTACGAACC CAGAACTGACAGAAG	18 15 17 17	TAACAG CAA CCACAG GTC CAACAG GCT CAACAG GAC	162 59 72 87
28	<i>cyt-2</i>	C D	1) CCTOGTATGT 2) TGC2GTAAGT or CTTTACTCACTATCT	CTGATCTAACCTCTT TTTTGCTAACGATGT CTTTACTCACTATCT	21 24 7	ATGTAG CCT ATCTAG CAC	92 95
29	<i>cyt-18</i>	S	1) TGG2GTAAGT	CATGACTAACACGAT	16	TACCAG CAC	64
30	<i>cyt-20</i>	C	1) ACG1GTTTCGT	ACCAACTTACACCTG	22	TTGTAG ACA	62
31	<i>cyt-21</i>	D	1) CAG2GTACGC or CTTOGTACGT 2) CTTOGTACGT	ACCAGCTAACTCTCT ACTCTCTGACTCCCA ATTGACTGACATTGC	19 11 25	TATTAG AAC AAACAG GTC	73 100
32	<i>for</i>	D	1) GCCOGTACGT 2) CAA2GTGAGT	GTTGACTCATAATCC CCAACTAATCCACC	16 x 17	ATACAG ATG CTCCAG GAT	80 63
33	<i>frq</i>	NO	INTRONS				
34	<i>grg-1</i>	D	1) AGG1GTAGGT 2) TCG2GTAAGC	AGACACTGACATCTC AGACACTAACATTCA	16 18	TCACAG GCG TCTCAG TCT	88 65
35	<i>H3</i>	C	1) CTCOGTAAGT	CGTTGCTAACGCGTC	14	ACCCAG GTC	67
35	<i>H4</i>	C	1) GAC1GTAAGT or CCA1GTACGT 2) CCA1GTACGT	GTGTTGTAACATCAT CATGACTGACTCGTA GTCAACTAACATGTC	29 17 17	CATCAG GCG CAACAG TGA	69 67
36	<i>his-3</i>	C	1) TGC1GTAAGT	TTTAACTCAAGACAC	9 x	ACATAG CTA	59
37	<i>hsp30</i>	NO	INTRONS				
38	<i>ilv-2</i>	D	1) ACG1GTGAGT 2) CTTOGTGAGT 3) AAGOGTGAGT 4) CGCOGTTCCG	GTTTACTGACGAGCT CCATGCTGACCCTTT GCGAGCTAACAAACA CCTGACTAACATTTG	19 18 15 18	ACACAG AGC CTTCAG GAC CAACAG ACC TCCTAG TCC	90 236 77 69
39	<i>laccase</i>	P	1) CGGOGTAAGT	AATGACTGACACACA	10	ACCTAG TAC	56
40	<i>leu-5</i>	S	1) TCGOGTGCGT	CCATGCTAATCCAGC	9	GCCCAG GAA	60
41	<i>leu-6</i>	D	1) CAGOGTACGC	CGGTACTAACTCGTC	11	TTCCAG GCC	63
42	<i>lox</i>	NO	INTRONS				
43	<i>met-7</i>	NO	INTRONS				
44	<i>mrp-3</i>	D	1) AACOGTAAGT	CCATGCTGACCATGC	23	CTCTAG CCC	307
45	<i>mta-1</i>	C	1) CAA2GTAAGT 2) TTT2GTAAGT	TTGTAAGTACCATTT TGAGACTAACCTCAC	12 9	CACTAG GAA ACTTAG CGG	53 57
46	<i>mtA-1</i>	D	1) GAT1GTGAGT	CATGGCTGATTGCTC	15 x	TTTCAG CGT	59
47	<i>nac</i>	S	1) GAG2GTGAGT 2) GACOGTACGT 3) GAGOGTATGT	TCAAATAACGGGTG CGTAACTGACCATTTG AGAGACTAAAAGTTCC	29 17 14 x	CTACAG CAG ACACAG GAC TTACAG ACA	234 691 64
48	<i>ncypt1</i>	D	1) TGA2GTAGTA 2) TTTOGTACGT 3) ATCOGTAAGC 4) AAGOGTACAC	CCCTGCTGACGATGC ACATGCTGACCGTTT ATTGCTGACCCAGT ACATACTTACACATC	11 11 16 14	AACCAG ATA GGCTAG AAA ATTGAG TGG CAACAG GAG	258 70 68 58
49	<i>nit-2</i>	D	1) CCG2GTATGT 2) CGCOGTGAGT	GCAAGCTAATTATAA CTGGTCTGATATATT	27 x 16 x	AAAAAG GAC GTCTAG AAC	99 78
50	<i>nit-3</i>	D	1) TTC1GTAAGT	TCCATCTAACTGACT	13	TCACAG ACA	61

Ref.	Gene	Id ^a	5' signal	Lariat signal	Dist ^b	3' signal	L ^c
51	<i>nit-4</i>	D	1)	GCA2GTAGGT AGTTACTCACCTTTT	8	TCACAG CAT	59
52	<i>nuc-1</i>		NO	INTRONS			
53	<i>nur22</i>	D	1)	TTA1GTACGA CTTGACTGACTTGTT	16	CAAAAG CTC	84
			2)	GAT1GTACGT AACAACTAATATTTT	17 x	TCACAG CCC	197
			3)	GGG1GTAGGT ACTTGCTAACCGAGC	19	ACAAAG GTA	81
54	<i>nur40</i>	D	1)	GCC1GTAATT GATTGCTAACACGTC	19	CCACAG GCT	66
			2)	TAG1GTACGA CATGGCTTATATCAA	17 x	TTGCAG CGA	71
			3)	ATTOGTGAGC AGCTACTAAGCATAA	17 x	CTCCAG GAG	93
			4)	CAA2GTGAGT TTGACCTGGGTCCCA	6 x	TCCCAG GAA	63
			5)	CAA2GTACGA GGTTCCTGATTGGAT	11 x	CTGTAG GGC	65
			6)	CACOGTCTTC AGCATCTGACAGCCG	11	TTTTAG GTG	59
			7)	CTT2GTAAAG GATGACTGATTCCCA	10 x	ATGCAG GCA	57
55	<i>nur49</i>	D	1)	ATGOGTGAGT TCAATCTAATATGTG	16 x	CCTTAG GAA	158
			2)	TTG1GTAAGT GGTAGCTAACCCTTT	20	TTCCAG GTG	84
56	<i>pho-4</i>	S	1)	TTG1GTATGT ATTCACTGACAACCA	21	CAACAG GAG	80
		D	2)	TGA1GTAAGT GCTTGCTAACGACGA	16	TTACAG AGC	83
57+58	<i>pma-1</i>	D	1)	GCTOGTAAGT GCATACTAACCCATT	11	GAATAG GAG	58
			2)	GAGOGTACGT CGATGCTGACTAGTT	14	CTACAG GGT	124
			3)	GAA2GTAGGT ACGCGCTAACCCGTT	15	TTTCAG GAC	64
			4)	TTG1GTAAGT AATAGCTAACAATAC	16	TCACAG TTG	67
59	<i>preg</i>	D	1)	CCG2GTAGGC AGCTGCTGACATGAA	14	TCATAG AAC	83
60	<i>pyr-4</i>		NO	INTRONS			
61	<i>qa-1F</i>		NO	INTRONS			
61+62	<i>qa-1S</i>	S	1)	TAG1GCACGT TCGTACTAACAGTCA	15	CACCAG GCT	66
61	<i>qa-2</i>		NO	INTRONS			
61	<i>qa-3</i>		NO	INTRONS			
61+63	<i>qa-4</i>		NO	INTRONS			
61	<i>qa-x</i>	S	1)	AAG2GTAAGT AAGTCCTGACACTGA	10	AAACAG CGC	69
			2)	ATG2GTGCGT GTTGACTAACAAGAA	19	GCTTAG GGA	74
61	<i>qa-y</i>		NO	INTRONS			
64	<i>sod-1</i>	P	1)	CTG1GTAAGC TACGGCTAACCTCTT	19	GTCCAG TCG	286
			2)	ACT1GTAAGT CTAGACTGACCAATG	24	CCGCAG TCA	100
			3)	GGCOGTATGT TCTTGCTAACTTTTA	11	CAACAG CGC	58
65	<i>spe-1</i>	D	1)	ATG1GTGAGT GTTTGCTGACTTGGA	18	CATCAG CCG	70
66	<i>T</i>	D	1)	ATG1GTGACC TTGTACTAACACAAA	12	ACCCAG GAG	52
			2)	TGT2GTATGT CGTCGCTGACAAGAA	20	CTGAAG TAA	99
67	<i>trp-1</i>		NO	INTRONS			
68	<i>trp-3</i>	C	1)	AGGOGTGCGT GCATGCTAACATCAC	18	CAACAG GCC	77
			2)	ACTOGTAAGA TCTTTCTGACACTTC	19	CTATAG ATT	72
69	<i>Ubi</i>	D	1)	CTT1GTAAGT CGATGCTAACTATCT	13	TCGCAG TGA	68
70	<i>ucr</i>	D	1)	AGG1GTGAGT GACAGCTGACGAGGC	20	ATACAG AGT	323
			2)	AACOCTAAGG GAATGCTGACCCCGG	14	TTACAG GTC	122
			3)	TTG1GTACGC GGAGACTGACATTTG	22	AAACAG GCG	101

Ref.	Gene	Id ^a	5' signal	Lariat signal	Dist ^b	3'signal	L ^c	
71	<i>vma-1</i>	D	1)	CCCOGTAAGC	GCCTGCTGACATGGC	15	GAATAG CAA	131
			2)	CCG1GTAAGT	TTATGCTAATAGCTC	9 x	TCGCAG GCA	74
			3)	CGG2GTGCGT	GCTCGCTAACCCATA	14	CCAAAG CCC	65
			4)	TTGOGTATGG	CTGAGCTGAGACTGG	11 x	AATTAG GTT	60
			5)	CGG1GTAAGG	GATGGCTAACCAATC	14	CGATAG CTG	63
			6)	AAGOGTATGT	TAAGGCTAACCATTT	18	CTATAG TAC	80
72	<i>vma-2</i>	D	1)	AATOGTTGGT	CATGTCTAACAACGG	11	CCGCAG GTC	56
			2)	GAG1GTGTGT	AACAGCTGACAGCCA	18	CTACAG GAA	71
			3)	CAGOGTAGAT	ATGCGCTGATATCAT	11 x	GAACAG GTC	59
			4)	AAGOGTGAGG	AGAACTGACCAGGA	12	CAACAG ACC	55
			5)	AGAOGTAAGT	TTGTGCTGACAAGAC	9	ACATAG GAA	58

Key: Ref. - Reference number for publication describing gene sequence, see list in accompanying paper.

a - Introns Identified by, C - computer analysis
D - cDNA sequencing
P - protein synthesis
S - Sl mapping

b - Distance between lariat consensus sequence and splice site of 3' consensus (bp).

c - Length of intron (bp).

x - Introns without a perfect CTNAC sequence within the Lariat consensus

The subscript number in the consensus sequences at the top of the table indicates the % occurrence of the particular nucleotide.

The number present within the 5' signal indicates the splicing position within the codon:

0 - does not cut codon

1 - cuts after 1st nucleotide within codon

2 - cuts after 2nd nucleotide within codon

Table IV Actual number of genes with a given number of introns compared with a Poisson Distribution

Number of introns n present in gene									
	0	1	2	3	4	5	6	7	8
Genes with n introns									
Expected Poisson Distribution	10	21	21	14	7	2.8	0.9	0.3	0.08
Observed Distribution	14	22	23	5	5	2	2	3	0