Volume 15 Number 4 1987 Nucleic Acids Research

Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates

Douglas R.Cavener

Department of Molecular Biology, Vanderbilt University, Nashville, TN 37235, USA

Received January 14, 1987; Accepted January 23, 1987

ABSTRACT

The previously presented consensus sequence for eukaryotic translation initiation sites by Kozak (1) was derived substantially from vertebrate mRNA sequences. Drosophila nuclear genes exhibit a significantly different translation start consensus sequence. These differences probably do not represent mechanistic differences in translation initiation inasmuch as both taxa exhibit identical preferences and restrictions at the crucial -3 position. Using more conservative criteria for the assignment of consensus the following consensus sequences were derived: vertebrate--CANCAUG and Drosophila--$^{C}_{A}$AA$^{A}_{C}$AUG.

INTRODUCTION

Previous analyses of the sequences flanking the translational start (TS) site in 211 eukaryotic mRNAs by Kozak (1) revealed an apparent consensus sequence CCACCAUG(G). Kozak's TS consensus sequence has been widely used to examine newly sequenced genes for the location of translational start sites. Kozak (2) has experimentally demonstrated that certain combinations of nucleotides flanking the start site have potent effects upon translation rates. This was most apparent at the -3 position (i.e. three nucleotides upstream of the start codon) where translation initiation is negatively affected by substitutions of nonconsensus nucleotides. The importance of the other consensus nucleotides is more subtle and they exhibit an interaction effect with the state of the -3 position. Sargan and coworkers (3) have proposed that the recognition of the start site by the ribosome could be mediated through complementary pairing of the mRNA CCACC sequence between -5 and -1 (or at least a similar sequence nearby) and five nucleotides at the base of the highly conserved 18S rRNA stem loop structure. Thus this consensus sequence has considerable practical and theoretical value. An untested assumption of this body of work is that the consensus sequence is valid for all eukaryotic taxa. Over 80% of the sequences analyzed by Kozak were of vertebrate origin and therefore the generality of this consensus sequence was unknown. I have compiled and analyzed the sequences flanking the start codon of Drosophila

Figure 1. Tabulated data and derived translation start consensus sequence.

VERTEBRATES

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5	+6
G	23	49	28	19	76	29	21	36	21	32		78	27	72
A	58	46	33	62	36	30	40	138	43	31		48	44	21
U	37	30	42	46	42	52	13	1	27	13		25	34	48
C	58	54	74	51	23	68	104	3	88	103		26	73	37
G	13	27	16	11	43	16	12	20	12	18		44	15	40
A	33	26	19	35	20	17	22	78	24	17		27	25	12
U	21	17	24	26	24	29	7	<1	15	7		14	19	27
C	33	30	42	29	13	38	58	2	49	58		15	41	21
	a/c	c	c	a	g	c	C	A	c	C	AUG	g	c	g

Vertebrate Consensus C A N C AUG

DROSOPHILA

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1		+4	+5	+6
G	13	16	14	10	19	15	2	10	7	14		18	11	18
A	29	29	23	35	29	21	25	63	43	29		25	23	7
U	9	19	17	13	14	22	6	1	8	6		15	10	15
C	24	11	21	17	15	19	44	2	19	28		10	24	28
G	17	21	19	13	25	19	3	13	9	18		26	16	26
A	39	39	31	47	38	27	32	82	56	38		37	34	10
U	12	25	23	17	18	29	8	1	10	8		22	15	22
C	32	15	28	23	19	25	57	3	25	36		15	35	41
	a	a	a	a	a	u	C/A	A	A	A/C	AUG	a	c	c

Drosophila Consensus C/A A A A/C AUG

For reference the ATG (AUG) start codon corresponds to +1 through +3. The vertebrate data was extracted from the compilation of sequences by Kozak (1). The Drosophila data are from the sequences listed in Fig. 2 with the exclusions indicated by asterisks. The first block for each data set contains the actual numerical data. The second block for each data set contains these same data presented as a percentage. Below the second block for each set is the derived consensus nucleotides (upper case letters) and preferred nucleotides (lower case letters) as defined in the text.

nuclear genes. In addition I have extracted the vertebrate data from Kozak's (1) compilation of sequences and analyzed them.

RESULTS AND DISCUSSION

Consensus criteria.

An important issue germane to the analysis of nucleic acid sequences is the criteria used for consensus assignments. In its common usage consensus means general agreement, quantitatively implying at least a majority. Thus it seems inappropriate to assign the status of consensus on the basis of a plurality of cases. With these considerations in mind I have chosen the following criteria for the assignment of consensus sequences. If the frequency of a single nucleotide at a specific position is greater than 50% and greater than twice the number of the second most frequent nucleotide it is assigned as the consensus nucleotide. If the sum of the frequencies of two nucleotides is greater than 75% (but neither meet the criteria for a single nucleotide assignment) they are assigned as co-consensus nucleotides. If no single nucleotide or pair of nucleotides meet the criteria of consensus nucleotide(s) the letter N is assigned to that position. (In such cases the most frequent nucleotide is denoted by a lower case letter in Figure 1).

The *in vivo* utilization of only a few of the start codons in the vertebrate and *Drosophila* data sets have been directly confirmed. Nonetheless, virtually all of the start codons in these two data sets have considerable indirect evidence supporting their identity. The type of evidence for the identity of the *Drosophila* start codons is indicated next to the sequences. I have not included several sequences for which an ambiguity occurs regarding the identification of the start codon. It is conceivable that a few of the start codons reported herein will eventually prove to be erroneous. However, the goal of this study was to obtain reliable consensus data which would not be significantly affected by a few errors.

Vertebrate TS consensus

The sequence data for all of the vertebrates were extracted from Kozak's compilation and analyzed (Figure 1). Not surprisingly, the consensus derived from these data generally agrees with the consensus derived by Kozak for the total data set (i.e. vertebrates and other higher eukaryotes). However, inspection of the numerical data indicate that there is no compelling consensus at the -5, -2, and +4 positions for vertebrates contrary to Kozak's

Figure 2 Sequences flanking Drosophila translation start codons.

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	123	456	Ref.	Start Data
Acetylcholinesterase	C	A	T	C	C	G	C	G	T	C	ATG	GCC	001	3
achaete-scute T5	A	T	C	T	C	T	T	A	A	A	ATG	GCT	002	2,3
Actin 79B	C	T	A	A	C	C	A	A	A	C	ATG	TGT	003	4
Actin 88F**	A	A	C	T	G	C	C	A	A	G	ATG	TGT	004	4
Alcohol dehydrogenase	A	G	A	A	G	T	C	A	C	C	ATG	TCG	005	1,4,5
Alcohol dehydrogenase (s)*	A	G	A	A	G	T	C	A	C	C	ATG	GCG	006	6
Alcohol dehydrogenase (m)*	A	G	A	A	G	T	C	A	C	C	ATG	GCG	007	6
Alcohol dehydrogenase (o)*	C	T	A	A	A	G	C	A	A	T	ATG	GCG	008	6
Alcohol dehydrogenase (p)*	A	A	A	A	G	A	C	A	G	A	ATG	TCT	009	6
Alcohol dehydrogenase (a)*	T	C	G	C	T	G	A	A	A	A	ATG	GTT	010	6
Alcohol dehydrogenase (h)*	C	A	C	A	G	A	A	A	A	A	ATG	GTT	011	6
Alcohol dehydrogenase-1 (mu)*	G	T	C	C	A	A	G	A	A	A	ATG	GCC	012	6
Alcohol dehydrogenase-2 (mu)*	C	T	C	C	A	T	T	G	A	A	ATG	GTT	013	6
3' gene to Adh	G	A	T	A	T	A	A	A	G	A	ATG	TTC	014	2
3' gene to Adh (s)*	G	A	T	A	G	A	A	A	G	A	ATG	TTC	015	2,6
3' gene to Adh (m)*	G	A	T	A	G	A	A	A	G	A	ATG	TTC	016	2,6
3' gene to Adh (p)*	A	G	C	C	A	A	A	A	G	A	ATG	TAC	017	2,6
Amylase	T	G	G	A	A	T	C	A	T	C	ATG	TTT	018	4
Amylase (p)*	C	T	A	G	C	A	T	A	A	C	ATG	TTC	019	6
Antennapedia	A	G	C	T	G	C	C	A	C	G	ATG	ACG	020	4,5
Aprt	A	A	G	T	A	G	A	A	A	A	ATG	AGC	021	3
bithoraxiod	A	C	T	T	G	A	A	A	T	A	ATG	AAT	022	2,4
bsg 25D	G	T	T	A	C	G	G	A	T	A	ATG	GAG	023	4
Calmodulin	A	C	C	T	A	C	A	A	A	A	ATG	GCC	024	1
Chorion s15-1	A	G	C	A	C	T	C	A	C	C	ATG	AAG	025	4
Chorion s18-1	C	A	G	C	C	T	C	A	G	A	ATG	ATG	026	4
Chorion s38-1	G	G	G	A	G	A	C	A	A	G	ATG	CAA	027	4
Chorion s36-1	A	A	A	C	G	G	C	A	A	C	ATG	ACG	028	3
Copia polyprotein	T	G	A	G	T	G	A	A	A	A	ATG	GAC	029	2
Cuticle protein I	G	T	C	A	G	C	C	A	A	T	ATG	TTC	030	2,6
Cuticle protein II*	A	T	C	A	G	C	C	A	A	C	ATG	TTC	031	2,6
Cuticle protein III**	C	C	A	A	A	T	C	A	A	A	ATG	TTC	032	2,6
Cuticle protein IV*	C	C	A	A	G	T	C	A	A	A	ATG	TTC	033	2,6
Dopa decarboxylase CNS	A	A	T	C	T	C	T	G	A	A	ATG	AGC	034	4
Dopa decarboxylase epidermal	C	A	A	G	A	T	C	G	A	C	ATG	GAG	035	4
Dras 1	C	C	A	C	A	G	C	C	A	A	ATG	ACG	036	2,6
Dras 2	C	A	G	T	C	T	T	A	T	A	ATG	TTT	037	3
Dsrc					T	A	A	G	C	C	ATG	GGC	038	2,6
Dsrc 28C	C	A	T	T	G	G	C	A	A	C	ATG	AAG	039	4,5
E74	C	C	T	A	T	C	A	G	C	G	ATG	CCC	040	4,5
EGF receptor homolog	T	G	A	G	C	A	C	A	T	C	ATG	AAT	041	2
engrailed	G	T	C	G	A	A	A	C	C	A	ATG	GCC	042	4,5
engrailed (v)*	A	A	G	T	G	A	A	C	A	A	ATG	GCC	043	3,6
Esterase-6	G	A	G	G	A	G	C	A	A	C	ATG	AAC	044	3
even-skipped	C	A	T	A	C	C	A	A	A	C	ATG	CAC	045	4
Gart	C	A	G	C	G	G	A	A	T	T	ATG	TCG	046	4
Glucose dehydrogenase	G	T	C	T	A	T	C	A	A	C	ATG	TCC	047	4
Hsp-70	C	T	C	A	C	A	C	A	C	A	ATG	CCT	048	4
Hsp-22	A	T	C	A	A	C	T	A	C	A	ATG	CGT	049	4
Hsp-23	A	A	A	A	A	C	A	A	A	A	ATG	GCA	050	4
Hsp-26	A	A	A	A	G	T	A	A	A	A	ATG	TCG	051	4
Hsp-27**	A	A	A	A	T	C	A	A	A	A	ATG	TCA	052	4
Hsp-82	T	A	C	A	T	A	C	A	A	G	ATG	CCA	053	4
Hsp-82 (s)*	T	A	A	A	T	A	C	A	A	G	ATG	CCA	054	4,6
Hsp-82 (p)*	C	A	C	A	T	A	C	A	A	G	ATG	CCC	055	4,6

	-10 -9 -8 -7 -6 -5 -4 -3 -2 -1	1 2 3	4 5 6	Ref.	Start Data
Hsp-82 (v)*	G A C A T A C A A G	ATG	CCT	056	4,6
LSP1 α	A G T T T C C A G G	ATG	AAG	057	4,6
LSP1 β**	A T C C G T C A A C	ATG	AAG	058	4,6
LSP1 γ**	A G G A C C A A G G	ATG	AAG	059	4,6
Mariner transposon ORF (m)	T G C A G T C A A C	ATG	TCG	060	2
Metallothionein	C T C A A T C A A G	ATG	CCT	061	4
Myosin light chain	A A C A G A C A A A	ATG	GCT	062	3
NHCP gene	A A A A A C A A A A	ATG	GGC	063	3
Opsin Rh2	G T A G C T G A G C	ATG	GAG	064	4
Opsin, ninaE**	C C A A A A C A C A	ATG	GAG	065	4
P-transposase	A T A A A A A A A A	ATG	AAA	066	4
paired	T C C A G A A A C T	ATG	ACC	067	2,3
period	C A G C A G C G A C	ATG	ATC	068	4
Polycomb	T T A A T T A A A A	ATG	ACT	069	3
Pupal cuticle gene	A C G C G A C A C C	ATG	TAT	070	2
Ribosomal protein A1	A G A C T T A A A C	ATG	CGT	071	3,4
Ribosomal protein 49	T T C A A G	ATG	ACC	072	4
RNA polymerase II, large sub.	G A C G A C C A G G	ATG	AGC	073	4
rosy, xanthine dehydrogenase	G C A C T T C A C G	ATG	TCT	074	3,5
rudimentary	C T C G T C C A A T	ATG	GCC	075	2,4,6
S60, 46C	C A G A A A A A A T	ATG	TCA	076	4
S72, 84B	C A T A C C A A A C	ATG	CAC	077	4
Sgs-3, glue protein	A G T A A A A A A C	ATG	AAG	078	4
Sgs-3 (s)*	A G T A A C A A A C	ATG	AAG	079	6
Sgs-3 (e)*	A G T A A C A A A C	ATG	AAG	080	6
Sgs-3 (y)*	A G T A A C A A A C	ATG	AAG	081	6
Sgs-4	C A A A G T C A A G	ATG	CGC	082	4
Sgs-5	C T T T T A C G A C	ATG	TTC	083	4
Sgs-7	A G A T A G A A C C	ATG	AAA	084	4
Sgs-8*	A G C A A C A A C C	ATG	AAG	085	4
Stellate	G T T C A A C C A G	ATG	GGC	086	2
sny β	C G G C G A C T A G	ATG	AGC	087	4
sry α	A T A G A A C A G C	ATG	GAA	088	4
sry γ	C G T C G G C G C A	ATG	GAT	089	4
Tropomyosin	C A C A A A C A C C	ATG	GAC	090	2
Tubulin, α1	A A A A C T C A A T	ATG	GTG	091	4,6
Tubulin, α2**	T T T G A T C A T C	ATG	GTA	092	4,6
Tubulin, α3*	A A A A A T C A A T	ATG	GCG	093	4,6
Tubulin, α4**	A A C T A A T A A A	ATG	GTG	094	4,6
Ultrabithorax	C A G C A G C G C A	ATG	AAC	095	4,5
Vitelline	A C C A A T C A A C	ATG	AAG	096	2,3
yellow	G C T A A G T G C A	ATG	TTC	097	4,5
Yolk protein-1	A A T C C G A A C C	ATG	AAC	098	4
Yolk protein-2**	G G A A G C C A C A	ATG	AAT	099	4
Yolk protein-3	T T G C A C C A A A	ATG	ATG	100	4

*Data not used for consensus analysis in Fig. 1. **Data used for analysis of positions -10 through -1 but not used for consensus analysis of positions +4 through +6. Above data is from D. melanogaster unless otherwise indicated by a letter abbreviation in parentheses. s = D. simulans, m = D. maritiana, o = D. orena, y = D. yakuba, e = D. erecta, p = D. pseudoobscura, v = D. virilis, mu = D. mulleri, a = D. affinidisjuncta, and h = D. hawaiiensis. Information used to identify the start codons is given in the Start Data column where 1 = Comparison of DNA sequence with amino acid sequence (independently determined), 2 = Open reading frame analysis of genomic DNA, 3 = Open reading frame analysis of cDNA, 4 = 5' transcript mapping data plus DNA sequence analysis, 5 = Analysis of in vitro transcription/translation products compared with DNA sequence, and 6 = Comparative analysis (interspecific or intraspecific) of homologous genes.

Bibliography for Figure 2. The citations are given in a condensed form. Personal communications are indicated by pc following the names.

001 Hall (1986) EMBO 5, 2949.
002 R. Villares & C. Cabrera, pc.
003 Fyrberg (1981) Cell 24, 107.
Sanchez (1983) JMB 163,533.
004 Fyrberg, op. cit.
Sanchez, op. cit.
005 Goldberg (1980) PNAS 77,5794.
Benyajati (1983) Cell 33,125.
006 Bodmer (1984) Nature 309,425
007 Bodmer, op. cit.
008 Bodmer, op. cit.
009 S. Schaeffer & C. Aquadro, pc.
010 R. Rowan & W. Dickinson, pc.
011 R. Rowan & W. Dickinson, pc.
012 Fischer (1985) NAR 13,6899.
013 Fischer, op. cit.
014 Cohn (1985) Ph.D. Thesis,
U. Michigan.
015 Cohn (1985) Ph.D. Thesis,
U. Michigan.
016 Cohn (1985) Ph.D. Thesis,
U. Michigan.
017 S. Schaeffer & C. Aquadro, pc.
018 Boer (1986) NAR 14, 8399.
019 C. Brown, pc.
020 Schneuwly (1986) EMBO 5,733.
Laughon (1986) MCB 6, 4676.
Stroeher (1986) MCB 6, 4667.
021 D. Johnson, pc.
022 H. Lipshitz, D. Peattie, & D.
Hogness, pc.
023 Boyer (1986) Ph.D. Thesis, UCLA
024 V. Smith, K. Doyle, J. Maune,
R.Munjaal & K. Beckingham, pc.
025 Levine (1985) Chromosoma
92,136.
026 Levine, op. cit.
027 B. Wakimoto, J. Levine, &
A. Spradling, pc.
028 J. Levine & A. Spradling, pc.
029 Mount (1985) MCB 5,1630.
030 Snyder (1982) Cell 29,1027.
031 Snyder, op. cit.
032 Snyder, op. cit.
033 Snyder, op. cit.
034 Morgan (1986) EMBO 5,3335.
Eveleth (1986) EMBO 5,2663.
035 Morgan, op. cit.
036 Neuman-Silberberg (1984)
Cell 37,1027.
037 Neuman-Silberg, op. cit.
H. Brock, pc.
038 Simon (1985) Cell 42,831.
039 S. Wadsworth, pc.
040 C. Thummel, K. Burtis,
& D. Hogness, pc.
041 Livneh (1985) Cell 40,599.
042 Poole (1985) Cell 40,37.
043 J. Kassis, D. Wright,
& P. O'Farrell, pc.
044 Nielsen, PNAS, in press.
045 Macdonald (1986) Cell 47, 721.
046 Henikoff (1986) Cell 44,33.
047 D. Cox, R. Whetten,
& D. Cavener, pc.
048 Torok (1980) NAR 8,3105.
Ingolia (1980) Cell 21,669.
049 Ingolia (1981) NAR 9,1627.
Southgate (1983) JMB 165,35.
050 Ingolia, op. cit.
Southgate, op. cit.
051 Ingolia, op. cit.
Southgate, op. cit.
052 Ingolia, op. cit.
Southgate, op. cit.
053 Blackman (1986) JMB 188,499.
054 Blackman, op. cit.
055 Blackman, op. cit.
056 Blackman, op. cit.
057 Delaney (1986) JMB 189,1.
Jowett (1986) EMBO 4,3789.
058 Delaney, op. cit.
059 Delaney, op. cit.
060 Jacobson (1986) PNAS 83,8684.
061 Maroni (1986) Genetics 112,493.
062 Falkenthal (1984) MCB 4,956.
063 James (1986) MCB 6, 3862.
064 Cowman (1986) Cell 44,705.
065 O'Tousa (1985) Cell 40,839.
Zuker (1985) Cell 40, 851.
066 Rio (1986) Cell 44,21.
O'Hare (1983) Cell 34,25.
067 Frigerio (1986) Cell 47, 735.
068 Jackson (1986) Nature 320,185.
069 R. Paro, pc.
070 Henikoff (1986) Cell 44,33.
071 S. Qian, pc.
072 O'Connell (1984) NAR 12,5495.
073 Biggs (1985) Cell 42,611.
074 C. Lee, D. Curtis, W. Bender,
M. McCarron, C. Love, &
A. Chovnick, pc.
075 Freund (1986) JMB 189, 25.
076 T. Hoey, pc.
077 T. Hoey, pc.
078 Garfinkel (1983) JMB 165,765.
079 C. Martin & E. Meyerowitz, pc.
080 C. Martin & E. Meyerowitz, pc.
081 C. Martin & E. Meyerowitz, pc.
082 Muskavitch (1982) Cell 29,1041.
083 Shore (1986) JMB 189,
084 Garfinkel (1983) JMB 165,765.

085 Garfinkel, op. cit.
086 K. Livak, pc.
087 Vincent (1985) JMB 186,149.
088 Vincent, op. cit.
089 Vincent, op. cit.
090 Karlik (1985) Cell 37,469.
091 Theurkauf (1986) PNAS 83,8477.
092 Theurkauf, op. cit.
093 Theurkauf, op. cit.
094 Theurkauf, op. cit.
095 R. Saint & H. Lipshitz, pc.
W. Petri & L. Scherer, pc.
096 W. Petri, p.c.
097 H. Biessmann, pc.
Chia (1986) EMBO 13,3597.
098 Hung (1981) NAR 9,6407.
099 Hung (1983) JMB 164,481.
100 Y. Yan, C. Kunert, & J. Postlethwait, pc.

consensus. Thus, the derived vertebrate consensus is CANCAUG using the consensus determination rules stated above.

Drosophila TS consensus

The sequence data for Drosophila were derived from published sequences and from unpublished reports sent to me (Figure 2). Data for the following six Drosophila genes were not included because of uncertainty of which start codon among multiple possibilities is actually used: white (4), zeste (V. Pirrotta, personal communication), fushi tarazu (5), a Hobo TE gene (R. Streck and S. Beckendorf, personal communication), Notch (6,7), caudal (P. Macdonald, personal communication and W. Gehring, personal communication) and Kruppel (8). The Drosophila data set contain a number of closely related genes. Gene sequences which were closely related to other genes in the data set were excluded from the consensus analysis and are not tabulated in Figure 1. The derived consensus for Drosophila is $\overset{C}{A}AA\overset{A}{C}AUG$. The average fit to the four consensus positions immediately upstream of the start codon is 3.1 nucleotides. Like vertebrates, Drosophila exhibits a strong consensus for A at the -3 position with a secondary preference for G. The major difference between the Drosophila and vertebrate consensus is that the Drosophila sequence is A biased as opposed to a C bias. Indeed, A is the most frequent nucleotide in 8 of 10 positions upstream of the start codon. This A bias yields differences between the Drosophila and vertebrate consensus at positions -4, -2, and -1. The G bias at the +4 position previously noted by Kozak (1) is not observed in Drosophila genes.

The differences between the vertebrate and Drosophila TS consensus sequences indicate that it is inappropriate to use the Kozak consensus as a general eukaryotic consensus sequence. These differences probably reflect taxonomic biases as opposed to qualitatively different mechanisms. Certainly one feature which may prove to be highly conserved in all higher eukaryotes is the strong preference for a purine at the -3 position. In addition C or A at positions -4, -2, and -1 may be a general preference. Finally the joint

occurrence of pyrimidines at the -3 and +4 positions is not observed in either data set. With the exception of this latter restriction, a wide range of sequence combinations is observed. Thus these consensus sequences cannot be used by themselves to discriminate between alternative start codons. However, the following summary of TS sequence frequencies for vertebrates and Drosophila may prove useful for the identification of putative start codons: RNNAUG 95-98%; YNNAUGR 2-5%; and YNNAUGY 0% (where R = purines and Y = pyrimidines).

Theoretical considerations.

Kozak (2) has provided compelling evidence in support of her scanning model of translation initiation. The scanning model proposes that ribosomes bind at the 5' cap of mRNAs and then scan (in a 5'-3' direction) for the first AUG in a good translation initiation context. An unresolved complication is that many mRNAs contain multiple AUGs in the "leader sequence" upstream of the start codon which initiates translation of the major coding region. In most cases these upstream AUGs are closely followed by stop codons. Kozak has demonstrated that such AUGs may be ignored by the ribosomes if they contain an exceptionally poor context (e.g. pyrimidines at -3 and +4). Although some of these upstream AUGs have a poor context others clearly have an adequate context as defined by the vertebrate and Drosophila consensus sequences defined herein. For example the Drosophila acetylcholinesterase (Ace) mRNA contains five upstream AUGs (9). Two of these are flanked by pyrimidines at -3 and +4. However the context of the other three AUGs fit the Drosophila consensus sequence just as well as the context of the start codon at the beginning of the 1,950 bp Ace coding region. Either these three short ORFs are translated as predicted by their context or they are ignored for some other reason (e.g. secondary structure exclusion). The Kruppel gene presents another type of complex sequence germane to translation initiation. The 5' end of the Kruppel mRNA contains four AUGs, all of which are flanked by -3/+4 pyrimidines (8). In contrast to Ace, these AUGs are not proceeded by stop codons and are in frame with the major reading frame. It is not known whether one of these AUGs serves as the start codon or whether the fifth AUG (which is in a good context) is the start codon.

The taxonomic differences reported herein are also relevant to molecular models which propose that the mRNA translation start site is recognized by the 18S ribosomal RNA (2,3). A highly conserved stem-loop structure exists at the 3' end of the 18S RNA (10). At the base of the stem is the sequence GGUGG which might base pair with the CCACC (-5 to -1) mRNA consensus sequence

proposed by Kozak. The former sequence is perfectly conserved in Drosophila, barley, and several vertebrates examined as well as several other eukaryotes. The data presented herein on the Drosophila TS consensus clearly present a difficult challenge to this model. The mean number of nucleotides in the Drosophila mRNA between -5 and -1 which are complementary to the 18S RNA GGUGG sequence is only 2.3 (+/- 1.0). It is possible that the GGUGG 18S sequence interacts with some other segment of the mRNA leader (3). However, it seems equally likely that interactions between the other elements of the ribosome and the sequences flanking the start codon are responsible for the proper localization of the start codon by the ribosome.

ACKNOWLEDGEMENTS

I would like to thank the Drosophila workers who shared unpublished sequences with me. I thank Diana Cox, Chris Schonbaum, Ross Whetten, Mike Murtha, Phil Krasney, Brian Foster, and Susan Schlitz for stimulating discussions concerning these data and for useful comments on the manuscript.

REFERENCES

1. Kozak, M. (1984) Nucleic Acids Res. 12, 857.
2. Kozak, M. (1986) Cell 44, 283-292.
3. Sargan, D.R., Gregory, S.P. and Butterworth, P.H.W. (1982) FEBS Letters 147, 133-136.
4. O'Hare, K., Murphy, C., Levis, R., and Rubin, G. (1984) J. Mol. Biol. 180, 437-455.
5. Laughon, A. and Scott, M. (1984) Nature 310, 25-31.
6. Wharton, K., Johansen, K., Xu, T. and Artavanis-Tsakonas, S. (1985) Cell 43, 567-581.
7. Kidd, S., Kelley, M. and Young, M. (1986) Mol. Cell. Biol. 6, 3094-3108.
8. Rosenberg, U., Schroder, C., Preiss, A., Cote, S., Riede, I., and Jackle, H. (1986) Nature 319, 336-339.
9. Hall, L. and Spierer, P. (1986) EMBO J. 5, 2949-2954.
10. Van Charldorp, R. and Van Knippenberg, P. (1982) Nucleic Acids Res. 10, 1149-1158.