Volume 12 Number 2 1984 Nucleic Acids Research

Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs

Marilyn Kozak

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received 15 November 1983; Accepted 21 November 1983

ABSTRACT

5'-Noncoding sequences have been tabulated for 211 messenger RNAs from higher eukaryotic cells. The 5'-proximal AUG triplet serves as the initiator codon in 95% of the mRNAs examined. The most conspicuous conserved feature is the presence of a purine (most often A) three nucleotides upstream from the AUG initiator codon; only 6 of the mRNAs in the survey have a pyrimidine in that position. There is a predominance of C in positions -1, -2, -4 and -5, just upstream from the initiator codon. The sequence CC$^{A}_{G}$CCAUG(G) thus emerges as a consensus sequence for eukaryotic initiation sites. The extent to which the ribosome binding site in a given mRNA matches the -1 to -5 consensus sequence varies: more than half of the mRNAs in the tabulation have 3 or 4 nucleotides in common with the CCACC consensus, but only ten mRNAs conform perfectly.

INTRODUCTION

Two years ago I prepared a compilation of the then-available 5'-noncoding sequences of eukaryotic mRNAs (Curr. Topics Microbiol. Immunol. 93, 81-123, 1981). Apart from a bevy of globin and histone mRNAs, only 32 other cellular mRNA sequences were known at that time. In contrast, there are 166 cellular mRNAs in the present compilation, not counting the globins and histones. I have excluded the mRNAs of lower eukaryotes and viruses, only to keep the survey manageable. One of my objectives was to determine whether certain patterns noted in the earlier compilation would be evident with this larger, more diversified set of sequences.

A few points about the selection and presentation of the sequences require explanation. In cases where numerous representatives of a gene family have been sequenced, I have omitted many and chosen those in which the leader sequences show the most divergence. There are exceptions, however. It seemed useful to include certain pairs of mRNAs in which the leader sequences show extensive homology *except near the AUG initiator codon* (e.g. human versus rat preproinsulin). The opposite pattern is also provocative; i.e., sequence conservation *only* near the AUG codon, as in human versus rat immunoglobulin E. Upon inspecting the completed compilation, only two families of mRNAs appeared to be excessively

represented: histones and globins. The 5'-noncoding sequences of histone mRNAs are sufficiently varied that they pose little danger of distorting the search for homology among ribosome binding sites. This is less true of the globin sequences, and I have controlled for this as described later in the text.

Nucleotide sequences determined by analyzing genomic DNA have been included only when there is sufficient supplementary data to identify introns that lie upstream from the AUG codon (or to verify their absence) and to estimate the location of the 5'-end of the mRNA. The 5'-end of each mRNA in the table was identified according to one of the following criteria:

(a) Direct sequence analysis of the purified mRNA.
(b) Primer-extension and/or mapping with a single-strand specific nuclease, such as S1. With these techniques there is often a 2- to 4-nucleotide ambiguity in pinpointing the cap site.
(c) Termination of the longest cDNA clone. When the cDNA clone is known to stop significantly short of the 5'-end of the mRNA, the sequence in the table is preceded by an ellipsis (...).
(d) Sequence homology with the corresponding gene from a closely-related species in which the 5'-terminus of the mRNA has been mapped.
(e) Presence in the genomic DNA sequence of an appropriately-positioned TATA box, 25- to 30-nucleotides upstream from the presumptive cap site. In the absence of other supporting data this criterion is rather weak.

The following criteria, identified by code letters in the rightmost column of the table, were used to identify the AUG initiator codon in each message:

(a) The nucleotide sequence corresponds to the known N-terminal amino acid sequence of the *primary* translation product. In some cases amino acid and nucleotide sequence data were derived from two different but related organisms.
(b) The N-terminal amino acid sequence has been determined only for the *mature* protein, which is known (or presumed) to derive from a precursor that carries an N-terminal extension (the "signal peptide") of 15 to 30 amino acids. The indicated AUG triplet is the only candidate initiation site compatible with the synthesis of such a precursor.
(c) The nucleotide sequence has a single open reading frame which either corresponds in size to the known molecular weight of the encoded protein, or includes peptides that are known to be present in the mature protein.
(d) The indicated AUG triplet occurs at the beginning of the longest open reading frame, but the exact size of the primary translation product is not available for comparison. This criterion is rather weak.
(e) The initiation site was deduced from sequence homology with the corresponding gene from a closely-related species in which the start site has been defined.
(f) Under conditions that allow formation of initiation complexes in vitro, the indicated AUG triplet was protected by ribosomes against nuclease digestion.

In 13 of the mRNAs in the table the functional initiator codon has not been definitively identified; the structure of the encoded protein is compatible with initiation at either of two nearby AUG triplets. In such cases I have *predicted* which AUG is most likely to be the (major) initiation site. Those entries are marked with an asterisk in the rightmost column. The AUG initiator codon was predicted based on position (i.e., proximity to the 5'-end of the mRNA) and conformity to

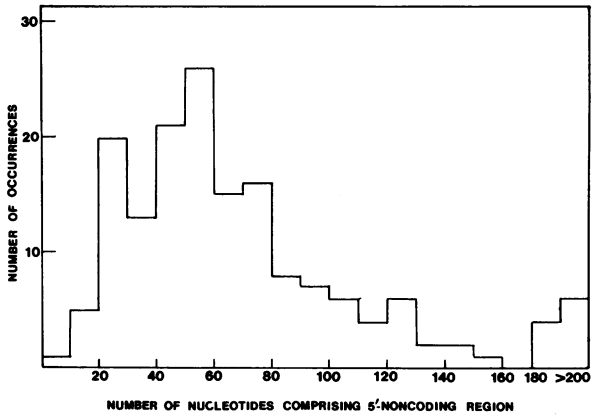

Figure 1. Length distribution of the 5'-noncoding portion of eukaryotic mRNAs. To avoid weighting the distribution by the large number of globin mRNAs in the sequence table, I scored the globin mRNAs only once in this tally.

the consensus sequence CCACCAUG. In a later section of the text I will explain in greater detail how this was done.

DISCUSSION

A few generalizations emerge from inspection of the sequences tabulated herein.

(i) The length of the 5'-noncoding region varies widely--from 3 to 572 nucleotides. However, 70% of the leader sequences are clustered in the 20- to 80-nucleotide range, as shown in Figure 1. The unusually long leader sequences occur on unusually interesting mRNAs (epidermal growth factor, oncogenes, heat shock proteins), inviting speculation that the structure of the 5'-noncoding region participates in the regulated expression of those genes.

(ii) Translation begins at the 5'-proximal AUG triplet in 95% of the mRNAs tabulated herein. There are only ten mRNAs listed in which one or more AUG triplets occur upstream from the recognized initiation site. The number of "nonfunctional" upstream AUG codons in each of those messages is shown in parentheses at the right edge of the table. [The upstream AUG's are called "nonfunctional" because there is as yet no evidence that ribosomes recognize those sites, but theory predicts that ribosomes should initiate (inefficiently) at the upstream AUG triplets as well as at the AUG codon that heads the long open reading frame.] The number of mRNAs with upstream AUG triplets would increase to 15 if my predictions are correct about which AUG is the major site of initiation in entries 80, 141, 146, 179 and 205. I have dealt elsewhere with the question of how ribosomes get past

ENTRY NO.[1]	MESSENGER RNA	SEQUENCES FROM THE 5'-NONCODING PORTION OF EUKARYOTIC CELLULAR mRNAs[2]	CAP SITE[3]	AUG[4]
001	Acetylcholine receptor pre-α-subunit (Torpedo)	[186]...GTTATTAGAAGTGGCAGATTTGCTTGAAAAGCCAATTATTGAAAGCTGAAGAATGATTCTG	c	b
002	Acetylcholine receptor pre-γ-subunit (Torpedo)	[119]...CCCTCACCCAACGACTCACTCACAAGCTGAGCTAGCACACTCACTGACGACCATGGTACTG	c	b (1)
003	Acetylcholine receptor pre-δ-subunit (Torpedo)	[456]...GACTGCAAATGCTATACTGCAACCAGTAAATTCGGACCTGAGCTCTTTCAAAATGGGGAAC	c	b (8)
004	α_1-Acid preGlycoprotein (rat)	...CTCTTCCTGGGCCGGTGCCTCTGAGTGTCTTCGGCATGGCGCTG	c	*
005	α-Actin, skeletal muscle (chicken)	[73]...GAGGCGTCTCTGCCAGCGGCCCGACGCGCAGTCAGCACAG/ACAGCCAGCAACATGTGTGAC	b	a
006	α-Actin, skeletal muscle (human)	[103]...GGCCCGAGCCGAGAGTAGCAG/TTGTAGCTACCCGCCCAGAAACTAGACACAATGTGCGAC	c	a
007	α-Actin, skeletal muscle (rat)	[71]...CTTCCTACCCTCGGCACCAGGGCCAGAGTCAGAGCAGCAG/AAACTAGACACCATGTGTGAC	b	a
008	β-Actin, cytoplasmic (rat)	[80]...ACAACCTCCTTGCAGCTCCTCCGTCCCGGTCCACACCCGCCACCAG/TTCGCCATGGATGAC	b	a
009	Actin, gene 79B (Drosophila)	[∿150]...CTGTTGTACCCTTGTACCCTTGTGTACCGCCCCGCACCAAACTAACCAAACATGTGTGAC	b	a
010	Actin, gene 88F (Drosophila)	[188]...TGGAGCTAACCGTGTGCACTTCCATCTCCCTTCCAGATAAACAACTGCCAAGATGTGTGAC	b	a (3)
011	Alcohol dehydrogenase adult form (Drosophila)	[123]...TACTTAATTGATCAA/ATCGAAAGAGCCTGCTAAAGCAAAAAAGAAGTCACCATGTCGTTT	b	a
012	pre-α-Amylase (barley)	[96]...AAGAAAAGGAGTGTTCTGTACTGTAAGTGAGACAATCGACAGTAGCGCGCGCCATGGGGAAG	c	c
013	pre-α-Amylase, pancreatic (mouse)	[17] GACAACTTCAAAGCAAAATGAAGTTC	a,b	a
014	pre-α-Amylase salivary -- (mouse) liver --	[95]...GCACATG/AAATAAATTAGTTGTTAGAAAGAATACTGCCAACAGCATAGCAAAATGAAATTC [205] CCAAATG/	b	a (1) (1)
015	preAngiotensinogen (rat)	[61]...GCTTGTCTGGGCTGGAGCTAAAGGACACACAGAAGCAAGTCCACAGATCCGTGATGACTCCC	c	b
016	preAntithrombin III (human)	[∿67] ...GATCACACTATCTCCACTTGCCCAGCCCTGTGGAAGATTAGCGGCCATGTATTCC	b	b,c
017	preproCalcitonin (rat)	[68]...CATCAGGACCCCGCAGTCTCAGCTCCAAGTCATCGCTCACCAG/GGAGGCATCATGGGCTTT	c	c
018	Calmodulin (chicken)	[91]...CGCACCGTGCAGCCGTTAGCTCCCACCGCACCCCGCAGCCGGGCCGAGCCACCATGGCTGAT	c	a
019	pre ε-Casein (mouse)	...TAGGAAGCAAGGACTCATTAATCATGAAGTTC	c	e

020	pre α-Casein (rat)	[61]...GATCATCTCCCAGCTTCTCTCACCCTACTCTTGGGTTCAAGATCTTAGCAACCA*TG*AAACTT	c	a	
021	pre β-Casein (rat)	[52] ATCCTCTGAGCTTCATCTTCTCTCTTGTCCTCCGCTAAAGGACTTGACAGCCA*TG*AAGGTC	b	a	
022	pre γ-Casein (rat)	[56]ATCATCATCTACCTATTCCTCTCGTCTTCCACTTGGGAAGCAAG$^{G}_{T}$ATCAAGTAACCA*TG*AAGTTC	b,e	a	
	preChorion proteins (silkworm)				
023	-ml911	...CTGAATATCCAGCATCA*TG*TTTACC	c	a,c	
024	-gene 292a, A-family	[35] ATCATTCTAGATTCAGCAGAACTGTTGTGATCAATA*TG*TCTACT	b	b	
025	-gene 18b, A-family	[35] GTCATTCTGAATTTAATATCTCATTCAGTGCAAACA*TG*TCTACC	b	b	
026	-gene 401a, B-family	[30] ATCATTCTCAGCTTTGATTTCAAAGTAAACA*TG*AACACT	b	b	
027	-gene 10a, B-family	[32] ATCATTATTGAGTTTCCTCTCATACAAACAAAA*TG*GCAGCC	b	b	
028	Chorionic gonadotropin pre-β-subunit (human)	...AGACAAGGCAGGGGACGCACCAAGGA*TG*GAGA*TG*	c	a	
029	preproChymosin (bovine)	[25] AGCAGCGGCTGGACCCAGATCCAAGA*TG*AGGTGT	c	b	
030	preproCollagen-α2, type I (chicken)	[133]...CCACGTCCGGGGGCTCTGCAACACAAGGAGTCTGCA*TG*TCTAGCAAGTAGACA*TG*CTCAGC	b	c (2)	
031	preComplement, C3 (mouse)	[56]GCCTCTGCCCACCCCTGCCCCTTACCCCTTCATTCCTTCCACCTTTTTCCTTCACTA*TG*GGACCA	b	b,c	
032	preConalbumin (chicken)	[76]...CTGTGACCAACACCGCTGCTCCCCCTCCTCAACACCCAGCTGCCTGCCCCAACA*TG*AAGCTC	a	a	
033	preproCorticotropin releasing factor (ovine)	[>127]...AGCGCTGCGCCTCGCTCACCTGCAGAAGCACCTCAGAAG/CGCCCCCTAACA*TG*CGACTG	c	d	
034	preCuticle protein I (Drosph)	[42] ATCAGTCAACGTTCGTTCTCGACCAGACAGAAGTCAGCCAATA*TG*TTCAAG	b	b	
035	preCuticle protein III "	[45] ATCAGTCTTAGAAGATTTCTAGTCCGACAATCCACCCAAATCAAAA*TG*TTCAAG	b	b	
036	Cytochrome P-450 (rat)	[30] ACTGAAGTCTACCGTGGTTACACCAGGACCA*TG*GAGCCC	b	a	
037	Dihydrofolate reductase (mouse)	[115]...TTGACGGCAATCCTAGCGTGAAGGCTGGTAGGATTTTATCCCCGCTGCCATCA*TG*GTTCGA (+ longer forms)	b	a	
038	preproElastase I (rat)	[21] GTGGTCTACTCTCTCCACAACA*TG*CTGCGC	b,c	b	
039	preproElastase II (rat)	[21] ACAGACATCCACGGACACACCA*TG*ATCAGG	b,c	c	
040	preproEnkephalin (human)	[129]...CTGAACCCGGCTTTTCCAATTGGCCTGCTCCATCCGAACAGCGTCAAC/TCCA*TG*GCGCGG	b	c	
041	preproEpidermal growth factor (mouse)	[353]...TCAGAGGCTCTCGAGAGGTGCAG$^{G}_{A}$AGGACCTGGAAAGGCAGCTAAATAAAAGA*TG*CCCTGG	b,c	d	
042	Fatty acid binding protein (rat liver)	[39] CTGTTGGTGGCAGCTGGGAAAGGAAACCTCATTGCCACCA*TG*AACTTC	c	a	
043	pre α-Fetoprotein (human)	[42] ATTGTGCTTCCACCACTGCCAATAACAAAATAACTAGCAACCA*TG*AAGTGG	b	c	
044	pre α-Fetoprotein (mouse)	[44] ACATCCCACTTCCAGCACTGCCTGCGGTGAAGGAAC$^{C}_{A}$AGCAGCCA*TG*AAGTGG	b,c	a	
045	preFibrinogen, Aα (human)	[>60]...TCCTTTCTTTCAGCTGGAGTGTCCTCAGGAGCCAGCCCCACCCTTAGAAAAGA*TG*TTTTCC	c	*	

046	preFibrinogen, γ (rat)	[53]	AGAGGTCACAGTGCTGGCTGTAAGGGGCTGGGAGCACCGGTCACCCAGACACTATGAATTGG	b,e	b
047	preFibroin (silkmoth)	[24]	ATCAGTTCGGTTCCAACTCTCAAGATGAGAGTC	a,b	c
048	preproGastrin (porcine)	[61]	...ACTGAGGCACCAGGCCAACAGCAGCACACCTGCCTCCCAGCTCTGCAGTCAAGATGCAGCGA	c	c (1)
	α-Globin family:				
049	-chicken, embryonic (π')	[55]	ACAACCTGCTCTGGGTGTTCACTGAAGGGAGCCTGAGCCAGCACTCTCCTGCACAATGGCACTG	d,e	a
050	-duck, major adult (α^A)	[36]	ACCCGTGCTGGGGGCTGCCAACGCGGAGCTGCAACCATGGTGCTG	b	a
051	-duck, minor adult (α^D)	[42]	ACAGAAACCCGTCAGTTGCCAGCCTGCCACGCCGCTGCCGCCATGCTGACC	b	a
052	-human, adult	[37]	ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTG	a	a
053	-human, embryonic (ζ)	[55]	ACCAAGGCCAGTCCTGAGCAGGCCCAACTCCAGTGCAGCTGCCCACCCTGCCGCCATGTCTCTG	b	a
054	-mouse, adult	[32]	ACTTCTGATTCTGACAGACTCAGGAAGAAACCATGGTGCTC	a	a
055	-rabbit, adult	[36]	ACACTTCTGGTCCAGTCCGACTGAGAAGGAACCACCATGGTGCTG	a	a
056	-Xenopus, major adult		...TGCACAACACAAACAGCAACCATGCTTCTT	c	c
	β-Globin family:				
057	-chicken, adult	[77]	...GAGCCCAGACCTCCTCCGTACCGACAACCACACGCTACCCTCCAACCGCCGCCATGGTGCAC	b	a
058	-duck, adult	[∿60]	...AGCCGAGACCTCCTCCGTACCTGCAGCCACACGCTACCCTCCGCCCGACACCATGGTGCAC	c	a
059	-chicken, embryonic (ρ)	[45]	AAGCTCTGAGTGCTCCCACAGCCGCACGCCAACCCCGCTGCCACCATGGTGCAC	c	a
060	-goat, adult (β^A)	[52]	ACACTTGCTTCTGACACAACCGTGTTCACTAGCAGCTACACAAACAGACACCATGCTGACT	d,e	a
061	-human, adult	[50]	ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCAC	a	a
062	-human, fetal (γ)	[53]	ACACTCGCTTCTGGAACGTCTGAG(A/G)TTATCAATAAGCTCCTAGTCCAGACGCCATGGGTCAT	c,d	a
063	-human, embryonic (ε)	[53]	ATATCTGCTTCCGACACAGCTGCAATCACTAGCAAGCTCTCAGGCCTGGCATCATGGTGCAT	d	a
064	-mouse, major adult	[52]	ACATTTGCTTCTGACATAGTTGTGTTGACTCACAACCCCAGAAACAGACATCATGGTGCAC	a	a
065	-mouse, embryonic	[52]	ACTTCGCTTCTGACACTCCTGTGATCACCAGCAACCTCCCAGACTTGCCATCATGGTGAAC	d	a
066	-rabbit, adult	[53]	ACACTTGCTTTTGACACAACTGTGTTTACTTGCAATCCCCCAAAACAGACAGAATGGTGCAT	a	a
067	-rabbit, embryonic (β3)	[62]	...TCTGAGACATCTGAGACTATCAGCAAGCTCAGCGAGCTCCTAGACCAGACATCATGGTTCAT	b	a
068	-Xenopus, major adult	[46]	ACTTGTTCTTTTTGCAGAAGCTCAGAATAAACGCTCAACTTTGGCCATGGGTTTG	b	c
069	-Xenopus, larval		...TAGCAGCTACTCCCATCTACAGCCACCATGGTGCAC	c	c
070	pre α_{2u} Globulin (rat)	[∿68]	...CCATCAGCAGAGAGATTGTCCCGACAGAGAGGCAATTCTATTCCCTACCAACATGAAGCTG	c	a
071	preproGlucagon I (anglerfish)	[∿58]	...AGGAACTAACAGCACTATTTGAGGGAGAAAAAGAATAAATACGGTTGTAAACATGAAACGC	c	c
072	preproGlucagon II (")		...GAAGCTCAAACAATGACAAGT	c	c
073	preproGlucagon (hamster)	[103]	...TGCACCTGCTCACCTGCTCTCCGCTCAGTCACAGCAGCAGGCAGAAAAAAAAATGAAGAAC	c	c
	preGlue proteins (Drosophila)				
074	-Sgs-4	[13]	TTCCAAAGTCAAGATGCGCTTG	b	d
075	-Sgs-3	[29]	ATCAGTTTGTGGAGAATTAAGTAAAAAACATGAAGCTG	c,d	d
076	-Sgs-7	[33]	ATCTGGTAAAGTAGTCTCAATCTAGATAGAACCATGAAACTG	b,d	d
077	-Sgs-8	[33]	ATCTGGTAAAGTTAATTTGTTAAAGCAACAACCATGAAGCTG	b,d	d
078	Glycoprotein hormones (human) pre-α-subunit	[100]	...CAGTCAACCGCCCTGAACACATCCTGCAAAAAGCCCAGAGAAAG/GAGCGCCATGGATTAC	d,e	a

079	Glycoprotein hormones (mouse) pre-α-subunit	[∿100]...AATCACCTGCCCAGACACATCCCTCAAAAGTCCAGAGCTTGCAGAAGAGCTATGGATTAC	c	b
080	preGrowth hormone (bovine)	[61]...AGGACCCAGTTCACCAGACGACTCAGGGTCCTGTGGACAGCTCACCAGCTATGATGGCTGCA/	d,e	*
081	preGrowth hormone (human)	[60]...AAGGCCCAACTCCCCGAACCACTCAGGGTCCTGTGGACGCTCACCTAGCTGCAATGGCTACA	b,e	b
082	preGrowth hormone (rat)	[60]...TCGAGCCCAGATTCCAAACTACTCAGGTCCTGTGGACAGATCACTGAGTGGCGATGGCTGCA	b,e	a,b
	Heat shock (Drosophila):			
083	-70K protein	[247]...GAGAACTCTGAATACTTTCAACAAGTCGTTACCGAGGAAGAACTCACACACAATGCCTGCT	b,e	c
084	-22K protein	[252]...ATTTGCATAAGAAGCTTTATTTGGAAAAACCCAAGTTACCTTATCAACTACAATGCGTTCC	b,e	c
085	-23K protein	[112]...AAGAAGTTTATTCTTTGAAGGAGAATCATCTTGAAGCAATTAAAAAACAAAAATGGCAAAT	b,e	c
086	-26K protein	[181]...AGAAAAAATATTTCAATCTCGCAAAGGAACATAACCTAAGGAAAAAGTAAAAATGTCGCTA	b,e	c
087	-27K protein	[121]...AAAAATTCCTTTGTCTAGACAGGGTTGTGAATAAAGAGAAAAAAAATCAAAAATGTCAATT	b,e	c
	Histocompatibility (MHC) antigens:			
088	-Class I: mouse (pre) $H\text{-}2K^d$	[25] AAGTCGCTAATCGCCGACCAGTGCGATGGCACCC	c,e	b
089	-Class II: mouse (pre) E_α 34K	[48] GCCTCAGTCTGCGATCGCTTCTGAACCCACCAAACACCCAAGAAGAAAATGGCCACA	b	b
090	- " mouse (pre) $I\text{-}A_\beta$ 29K	...GGCATTACCTGTGCCTTAGAGATGGCTCTG	-	b
091	- " human (pre) DR_β 29K	...CTCCTCTGGCCCCTGGTCCTGTCCTCTTCTCCAGCATGGTGTGT	c	b
	Histones:			
092	-chicken H1, embryonic	[38] AGTGGTTCCCCCCGATCTGTGGAACGACGTCCGTCACCATGTCGGAG	b	c
093	- " H2A.1, adult	[146]...GTTGCGCGCTGCGTTCTTGGCTTGTTTGCTCTCTGAGTGTTCAGTCGCTGCGATGTCGGGG	e	a
094	- " H2A.F, embryonic	[70]...GAGGCGGTTCGAGCGGCGGATTCGGGGACCGGGGCTCGGGCGGCGGCGGCACCATGGCAGGT	c	c
095	- " H2B	[36] ACTCCGCTGCGCCGAAGGGATCGTGGAGAGTTCGACATGCCTGAG	e	a
096	- " H4, embryonic	[26] GGTGTCGGACGACTCAGGCTCTCGGCATGTCTGGC	b	a
097	- " H5, erythrocyte	[109]...CTTTTTTAAGCTCCCTAACCCCAGTGCCCTGCCGTGGGGTGAAGCGGCGGCCATGACGGAG	c,e	a
098	-human H3	[27] ACTTTCGGTTGCAGAGCAGTTCTGCGAATGGCGCGT	e	a
099	- " H4	[36] ACAAAGCGTTGGGTGAGACGCCTCTTGCTCGTCGTCATGTCTGGC	e	a
100	-sea urchin, early, H1	[38] ATTCGTTTTGTTAACCTCCCGACGCACCGTATATCAAGATGGCTGAG	b	
101	(S. purpuratus) H2A	[70]...TTGTTACGTTCTGAACTTCGCTTCCGATTTATTCTAAACTCATCAACAACATCATGTCTGGC	b	a
102	H2B	[78]...CTTGACATACTCGTTTCGCTGCATCTTTACAGACCAGAAAACCTCAATTCATCATGGCTCCA	b	a
103	H3	[55] ATTCATCCCGTCACTCGTATTTGAAGTACTGAATTACCTACTGTCCCAAGCAACTATGGCACGC	b	a
104	H4	[67]...ATATCCAGTCTACAGGATCACACAGAACTCGCTCTCAACTATCAATCATCATCATGTCAGGT	a,b	a
105	-sea urchin, early H2A	[60] ...CGCTTCGTTCACAACCTCGCTTCGCTCTCAGCTCGTTAACCAACCAACCATCATGTCTGGA	b	a
106	(P. miliaris, h22) H2B	[76]...TCGTGACACTCGCATCGTTCTGCTCCTAAGACATCAGAAAACTTCATCTCACCATG	b	a
107	H3	[54] ATTCATCTCGTCACCCTCGTTTGAAACACTGATCCAACTGTCTCCCAATCAACCATGGCACGA	b	a
108	H4	[60]...AGCGAAACGTCCAGTCGTCAGCATCGCACTAAGACTCTCTCTCAATCTCCATAATGTCAGGC	d	a
109	-sea urchin, late H3	[∿27] ATCAGTTGACTATCTGAAAAATCAACAATGGCCCGT	b,e	a
110	(L. pictus) H4	[∿23] ATCTCAAACAGGAAACTATCAACATGTCTGGA	b,e	a
111	-Xenopus H1	[28] AGTTGGCGTGAGTGAATTTACTTCAAAGATGACAGAA	e	a
112	" H2A	[47] AGTCTACAACATCTTCTTGATTGTGGTTGATTTGTAGCACAGTAATCATGTCTGGA	e	a
113	" H2B	[35] ACAGTTTTGTAGGCTGAGAGAGAAGCAGCACAATTATGCCTGAA	e	a
114	" H3	[80?]...CGGGTTACCCGCTGCCGCAGTTACACGCTTAGGTTTCTTTAACTGATACACTATGGCCCGT	e	a
115	" H4	[28] ATATTGTGTTTCAAGAGCTCAAGAAAGAATGTCTGGA	e	a

116	Hypoxanthine phosphoribosyl-transferase (HPRT), human	[85]...GCCACCGGCTTCCTCCTCCTGAGCAGTCAGCCCGCGCGCCGGCCGGCTCCGTT*ATG*GCGACC	c	c	
117	HPRT, mouse	[87]...TCCGCCGGCTTCCTCCTCAGACCGCTTTTTGCCGCGAGCCGACCGGTCCCGTC*ATG*CCGACC	c	c	
	Immunoglobulins:				
118	-(pre)κ_{167} light chain, mouse	[29] ATCAGAACTCTTCACACCAAGTTCTCAGA*ATG*AGGTGC	b	b	
119	-(pre)κ light chain, mouse	[3; 18] GAA*ATG*CATCACACCAGC*ATG*GGCATC	a,b	a^+	
120	-(pre)λ_I light chain, mouse	[40] GCTGACCAATATTGAAAAGAATAGACCTGGTTTGTGAATT*ATG*GCCTGG	b	a	
121	-(pre)heavy chain, mouse	[45] TGTCCCAATCTTCACATTCAGAAATCAGCACTCAGTCCTGTCAC*ATG*AAGTTG	c	b	
122	-(pre)heavy chain, mouse	[30] ATCCTCTTCTCATAGAGCCTCCGTCAGAGC*ATG*GCTGTC	b	c	
123	-(pre)heavy chain, human IgE	[56]CACCCAGCAACCACATCTGTCCTCTAGAGAATCCCCTGAGAGCTCCGTTCCTCACC*ATG*GACTGG	c	b	
124	-(pre)heavy chain, rat IgE	[71]...TGAACGCACAAATCTTTTCCTTCAGTGACTAGCACTGATACAGAATATTCACC*ATG*GACTTG	c	b	
125	preproInsulin (chicken)	[54] AGCCTG*ATG*AATAAAATATTCCTTTCCTCTTCAGAAG/GCCTCCCCCAGCTCATC*ATG*GCTCTC	d	b (1)	
126	preproInsulin (human)	[59]...CAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAG/ATCACTGTCCTTCTGCC*ATG*GCCCTG	d	b	
127	preproInsulin (rat-I)	[57]...TAAGTGACCAGCTACAATCATAGACCATCAGCAAGCAG/GTCATTGTTCCAAC*ATG*GCCCTG	b,e	a	
128	preproInsulin (anglerfish)	[85]...CTCTTCATCAACTCATTCATCACCTCCTCCTACTCTACAG/TTCTACTGCAGC*ATG*GCGGCT	c	a	
129	preproInsulin (hagfish)	[49] CCAACTTGTCGCACTAGAGGGACGGAGACCATTATCCAAAAGCAACCAA*ATG*GCGCTC	c	b	
130	preproInsulin (salmon)	[>71]...ATTCTTCTACAGAGCTTCTTCACCACCACCAACTCCATAGCCTACCATCACC*ATG*GCCCTC	c	b	
	Interferons:				
131	-pre-α (LeIF-A), human	[67]...AAGGTTTAGGCTCACCCATTTCAACCAGTCTAGCAGCATCTGCAACATCTACA*ATG*GCCTTG	c,e	b	
132	-pre-α (LeIF-D), human	[67]...AAGGTTCAGAGTCACCCATCTCAGCAAGCCCAGAAGTATCTGCAATATCTACG*ATG*GCCTCG	c,e	b	
133	-pre-α (LeIF-C), human	[67]...CAAAGTTCAAGGTTACCCATCTCAAGTAGCCTAGCAACATTTGCAACATCCCAA*ATG*GCCCTG	b	b	
134	-pre-α (LeIF-H), human	[69]...AGGTTCAGTGTTACCCCTCATCAACCAGCCCAGCAGCATCTTCAGGGTTCCCAA*ATG*GCATTG	b	b	
135	-pre-α (leukocyte), mouse	[67]...GAAGGATCAGGACCAAACAGTCCAGAAGACCAGAAGCTTTGGCAACACTCACC*ATG*GCTAGG	d,e	b	
136	-pre-β (fibroblast), human	[75]...AAGCCTTTGCTCTGGCACAACAGGTAGTAGGCGACACTGTTCGTGTTGTCAAC*ATG*ACCAAC	b,c	b	
137	-pre-β (fibroblast), mouse	...ACACCAGCCTGGCTTCCATC*ATG*AACAAC	c	b	
138	-pre-γ (immune), human	[128]...TGATACAAGAACTACTGATTTCAACTTCTTTGGCTTAATTCTCTCGGAAACG*ATG*AAATAT	b,e	d	
139	preInterleukin-2, human	[47] ATCACTCTCTTTAATCACTACTCACAGTAACCTCAACTCCTGCCACA*ATG*TACAGG	b,c	*	
140	preproKallikrein, mouse	[42] AGCTCCAAGCTCACTGCCTGCTGCTCCTGAACACCTGTTACC*ATG*TGGTTC	b	b	
141	preproKallikrein, rat	...AAGCTCAGCACCTGCTGCTCCTGC*ATG*CCTGTTACC*ATG*TGGTTC	c	*	
142	Keratin, epidermal (mouse)	...CCGACTCCTGGGCCACACCACACCACC*ATG*TCTGTC	c	b	
143	Keratin, B2A (sheep)	[50] ACTCAAAAACTTCTCTTAACAACCCTCCTCTCAACCCAACTCCTGACACC*ATG*GCCTGC	e	a	
144	preKininogen, low MW (bovine)	[133]...CCCAACTGGCTCCCAATTCCTGGTGAAACTCCTAAAGGGAGACTGTTAAATC*ATG*AAATTA	c	c	
145	pre-α-Lactalbumin, guinea pig	[84]...TTCTGGGTGT*ATG*GGAAGAGCAGGGGCATTTCAGGTTCACAGCAGCAGCCAAAA*ATGATG*TCC	c	b (2)	
146	preLectin (French bean)	[16] NN*ATG*AA*TG*C*ATG*ATC*ATG*GCTTCC	c	*	
147	Leghemoglobin (soybean)	[49] AACTTGCATTGAACAATTAATAGAAATAACAGAAAAGTAGAAAAGAAAT*ATG*GGTGCT	b,e	a	

148	apoLipoprotein II, VLD, chick	[77]	...GAAAGCAGGACAG/GTCTCTTGGTGTAAAGGGCTGAACTGGTACCAACAAACC*ATG*GTGCAA	b	a
149	apoLipoprotein A-I (human)		...TCCCCCACGGCCCTTCAGG*ATG*AAAGCT	c	a
150	apoLipoprotein E (rat)		...ACTGGCCAATCACAACTGGGAAG*ATG*AAGGCT	c	a
151	preLysozyme (chicken)	[29]	AGTCCCGCTGTGTGTACGACACTGGCAAC*ATG*AGGTCT	b	a
152	preproMelittin (honeybee)		...AGCGAATTAACAGAATTAACAGGAAGGAAGGAAGGAAGCGATCGGAGAAATC*ATG*AAATTC	c	a
153	Metallothionein-II (human)	[69]	...CCAGCGAACCCGCGTGCAACCTGTCCCGACTCTAGCCGCCTCTTCAGCTCGCC*ATG*GATCCC	b	a
154	Metallothionein-I (mouse)	[73]	...CTGAGTACCTTCTCCTCACTTACTCCGTAGCTCCAGCTTCACCAGATCTCGG*AATG*GACCCC	b	a
155	pre-β2-Microglobulin (mouse)	[>52]	...ATTTCAGTGGCTGCTACTCGGCGCTTCAGTCGCGGTCGCTTCAGTCGTCAGC*ATG*GCTCGC	c	a
156	Myoglobin (seal)	[70]	...CAGGACACCGAGTCAGCCGGGACTTGCTCTTCTTGTCTTCCTCAGACTGCACC*ATG*GGGCTC	b	a
157	Myosin, skel. L chain (chick)		...CCTCTCAGCTAATCCCTCCGGCCCGTCCGCTCTACTTTTCCAACTCTCAATC*ATG*TCCTTC	c	c
158	prepro-β-Nerve growth factor (mouse)	[74]	...GCTGGCCTTATATTTGGATCTCCCGGGCAGCTTTTTGGAAACTCCTAGTGAAC*ATG*CTGTGC	c	d
	protoOncogenes:				
159	-c-*fos* (human)	[154]	...CCCACCTGTCTCCGCCCCTCGGCCCCTCGCCCGGCTTTGCCTAACCGCCACG*ATGATG*TTC	e	*
160	-c-*myc* (human)	[572]	...AGACGCTGGATTTTTTTCGGGTAGTGGAAAACCAG/CAGCCTCCCGCGACG*ATG*CCCCTC	c	d
161	-c-Ki-*ras*2 (human)	[>181]	...GGGGCCAGAGGCTCAGCGGCTCCCAGGTGCGGGAGAGAG/GCCTGCTGAAAA*TG*ACTGAA	c	c
162	-c-*src* (chicken)		...TCGCGTACCACTGTGGCCAGGCGGTAGCTGGGACGTGCAG/CCCACCACC*ATG*GGGAGC	-	c
163	preproOpiomelanocortin (bovine)	[128]	...GGAGCCGCCCGAGGCAGCTTCCCCGTGACAG/AGCCTCAGCCTGCCTGGAAG*ATG*CCGAGA	b	a
164	preproOpiomelanocortin (human)	[107]	...CCCGCCCTCAGAGAGCAGCCTCCCGAGACAG/AGCCTCAGCCTGCCTGGAAG*ATG*CCGAGA	d,e	a
165	Ovalbumin (chicken)	[64]	...AAAGCTGTATTGCCTTTAGCAGTCAAGCTCGAAAG/ACAACTCAGAGTTCACC*ATG*GGCTCC	a,b	a,f
166	preOvomucoid (chicken)	[53]	ATCTCAGGAGCAGAGCACCGGCAGCCGCCTGCAGAGCCGGGCAGTACCTCACC*ATG*GCC*ATG*	b,c	a
167	preproParathyroid hormone (bovine)	[100]	...TCAGCTGCTAATACATTTGAAAGAAGATTGTATCCTAAGACGTGTG/TTAAT*ATGATG*TCT	b	a
168	preproParathyroid hormone (human)	[90]	...AGCTACTAACATACCTGAACGAAGATCTTGTTCTAAGACATTGT*ATG*/TGAAG*ATG*ATACCT	d,e	a (1)
169	prePepsinogen (human)	[∿54]	CTGCACCTTCCTCCCGTCTTGCCTTCTCCCTCGAGTTGGGACCCGGGAAGAACC*ATG*AAGTGG	b,e	b
170	pre(?)Phaseolin	[77]	...CTACTCTACTACTATAATACCCCAACCCAACTCATATTCAATACTACTCTACT*ATGATG*AGA	c	*
171	Phosphoglycerate kinase, human	[80]	...GGCTCCCTCGTTGACCGAATCACCGACCTCTCTCCCCAGCTGTATTTCCAAAA*TG*TCGCTT	c	a
172	prePlacental lactogen, human		...CTGTGGACAGCTCACCTAGCGGCAA*TG*GCTGCA	c	a

173	prePlasminogen activator (human)	[84]...GCGAGGAAAGGGAAGGAGCAAGCCGTGAATTTAAGGGACGCTGTGAAGCAATC**ATGGATGC**A	c	*
174	preProlactin (rat)	[51] AGTGGTTCTCTTAGGACTTCTTGGGGAAGTGTGGTCCCAGTGGTCATCACC**ATG**AACAGC	c,e	a
175	preProlactin (bovine)	[67]...GCCATAGGACGAGAGCTTCCTGGTGAAGTGTGTTTCTTGAAATCATCACCACC**ATG**GACAGC	c	b
176	Protamine (trout)	[14] ATCCATCAATCACA**ATG**CCCAGA	b	a
177	Pyruvate kinase (chicken)	[80]...GCTTTGGGCACGGCGGCGGAAGCAGCAGCAGGAGACACCGAACTCCAGTAACC**ATG**TCGAAG	c	c
178	preproRelaxin (rat)	[∿60] ...CTGAACCGCCCAGGAGCACCGCCCAGGAGCAGCCCAGACCGGA**ATG**TCCAGC	c	b
179	preproRenin (mouse)	[55] TCTGGGCTACACAGCTCTTAGAAAGCCTTGGCTGAACCAG**ATG**GACAGGAGGAGG**ATG**CCTCTC	c	*
180	preRibonuclease, panc. (rat)	[∿75]...CAATTTGCTCTGGAATTCAAACGTTTAGGCTCCTCAGAGCAGCAAAGCCACT**ATG**GGTCTG	c	b
181	Ribosomal protein S19, Xenopus	[>46] ...AATTCTTTAATCCTTTCTTTTGTCACCGTGAGAGATAGCCGGCAAG**ATG**AATGAC	c	c
182	preRibulose bisphosphate carboxylase (soybean)	[45] ATCTGGCAGCAGAAAAACAAGTAGTTGAGAACTAAGAAGAAGAAAA**ATG**GCTTCC	b	a
183	Seminal vesicle pre-secretory protein IV (rat)	[22] AGTCAAGAGCTTTTCTGGCAAG**ATG**AAGTCT	b	b
184	pre(?)Sericin (silkmoth)	[54] ATAGTCGTCTTATCATCGGGTCTCTAAGGATCAAGCGATCCAAAGACCGCCAAC**ATG**CGTTTC	b	c
185	Serum preproalbumin, chick	[41] AGCATTTTTGAATAATTTAGCCCACATCATAATCTGCAGCC**ATG**AAGTGG	b	a
186	Serum preproalbumin, human	...GCTTTTCTCTTCTGTCAACCCCACACGCCTTTGGCACA**ATG**AAGTGG	c	a
187	preproSomatocrinin (human)	[112]...TGCCTCGGAGCAGAGGGATATCTGCCGCATCAGGTGCCACCCCGGGTGAAGG**ATG**CCACTC	b	c
188	preproSomatostatin-II (anglerfish)	[59]...CAAACCCAGCAGAACCAGTAGAACCAGCAGAAGACACCAGACCAGCAGACAGT**ATG**CAGTGT	c	c
189	preproSomatostatin-22 (catfish)	...CCGCCCAGCTCGAAATCTTCACCAGCTACCAAGAAG**ATG**TCGTCT	c	c
190	preproSomatostatin-14 "	[114]...GCCCTCCCTCCACCAATTTTTCCACCAAATCCAGCTTTATTTCTTTTTGAAG**ATG**CCCTCC	c	c
191	preproSomatostatin-I (human)	[105]...CGCGCCTAGAGTTTGACCAGCCACTCTCCAGCTCGGCTTTCGCGGCGCCGAG**ATG**CTGTCC	c	c
192	preproSomatostatin (rat)	[81]...CTGCGTCTAGACTGACCCACCGCGCTCAAGCTCGGCTGTCTGAGGCAGGGGAG**ATG**CTGTCC	c	c
193	pre-Steroid binding protein C1, rat prostate	[44] TGGAAGAGTTCATTTGCTCACCATTTGCTAAGTAGAAAACTGAA**ATG**AGCACC	b	b
194	pre-Steroid binding protein C2, rat prostate	[42] TGGGAGAGTCCATTTGCTCAGTCTAAAAGCAAACTGAGCACC**ATG**AGGCTG	b	b
195	pre-Steroid binding protein C3, rat prostate	[55] GAGTTTCCTGATTTCTGTCTTGGACAACAGAACAACCCACAGGGACTGCCTCAAC**ATG**AAGCTG	b	b
196	preproThaumatin	[31] AAAGCGCAGCCTCAATTGGCATCATACATCAA**ATG**GCCGCC	b	c
197	Thyrotropin, pre-β-subunit (rat)	[89]...CGAGTGGAGAGAAAAATATTCTGCTTCAGTGAAGAGCTGGGGTTGTTCAAAGC**ATG**AGTGCT	c	b
198	preproTrypsinogen (rat)	[12] CCTTCTGCCACC**ATG**AGTGCA	b	b

199	α-Tubulin (rat)	[100]...AACACCTCCTCCTCGCCTCCGCCATCCACCCGGCAGCCGCGAAGCAGCAACC*ATG*/CGTGA	b	a
200	β-Tubulin (chicken)	[87]...AGAGCGGGAGGTGACGGAGCGGGAGCAGCGCGGCACCGGCAGACACCGGCATC*ATG*CGTGAG	c	a
201	β-Tubulin (human)	[159]...TTTTCTTGCCCCATACATACCTTGAGGCGAGCAAAAAAATTAAATTTTAACC*ATG*AGGGAA	b	a
202	preUteroglobin (rabbit)	[47] AGATCACCGGATCCAGAGCCAGCCCAGAGCCTTCCCATTCTGCCACC*ATG*AAGCTC	b,c	a
203	preproVasoactive intestinal polypeptide (human)	GGGGAGCACGACTGGGCGAGAGGCACAGAA*ATG*GACACC	b	c
204	preproVasopressin neurophysin II (bovine)	[49] GCACAGTCTACAGAGCAGCACTGCGCACGTGTGCCCACGCGTGCCAGG*ATG*CCCGAC	b	b
205	preproVasopressin neurophysin (rat)	AGCAGAGCAGAGCTGCACGCAGTGCCCACCT*ATG*CTCGCC*ATGATG*CTCAAC	b	*
206	Vitellogenin II (chick)	[13] ATTCACCTTCGCT*ATG*AGGGGG	b	d
207	Vitellogenin I (Drosophila)	[≥58]...ACTCACTCAGTGTTGAAGTCGCATCCGCAGGACCAAATCCCAAATCCGAACC*ATG*AACCCC	b	*
208	Vitellogenin II (Drosophila)	[≥51] *ATG*CAGTACAATTTGGTACGGTGTCTGAAAAAGTCGAACTTGGAAGCCACA*ATG*AATCCT	b	*
209	preWhey acidic protein (mouse)	[26] ATCAGTCACTTGCCTGACCCGGTACC*ATG*CGTTGC		c
210	preZein, 19K (maize)	[57]...CACATATTATTGAGACCAACTAGCAATATAGAAAGCACAATATTGTACCAATA*ATG*GCAGCC	c	b
211	preZein, 22K (maize)	[>67]...TCAGCATTCAAAAACACACCAAGCGAAGCGCACTAGCAAC$^{G}_{A}$ACCTAACA$^{A}_{C}$C*AATG*GCTACC	c	c

Footnotes:

[1]The number assigned here to each mRNA is used again to identify the corresponding references in the bibliography.

[2]The table shows the sequence of the plus strand of DNA, from which the sequence of mRNA can be derived by substituting U for T. All ATG triplets are shown in italics. The sequences are aligned by using the ATG triplet that is known (see footnote 4) or predicted (see text) to be the functional initiator codon. The positions of introns are indicated by a diagonal line. For mRNAs in which the 5'-noncoding sequence exceeds 56 nucleotides, I have shown only the portion nearest the ATG triplet. *The number in brackets preceding the sequence indicates the (approximate) full length of the 5'-noncoding sequence,* not counting the m^7G cap or the ATG triplet. There is likely to be a 2- to 4-nucleotide uncertainty when S1 nuclease was used to map the cap site. In cases where cDNA clones were analyzed, the 5'-noncoding sequence may be a little longer than indicated. If the missing portion of the leader is suspected to be more than a few nucleotides, no figure is given for the overall length of the 5'-noncoding region.

[3]The criteria (a-e) used to identify the 5'-terminus of each mRNA are summarized in the text.

[4]The criteria (a-f) used to identify the ATG initiator codon are summarized in the text. An *asterisk* in this column means that the ATG triplet used to align the mRNA is *predicted,* but is not known to be the functional initiator codon. I have also marked in this column those mRNAs that have ATG triplets upstream from the functional initiation site. *The number of upstream ATG's is indicated in parentheses.* Entry 119 (marked +) is, to my knowledge, the only cellular mRNA in which two functional initiator codons have been identified: the ATGs in positions 4-6, and 19-21.

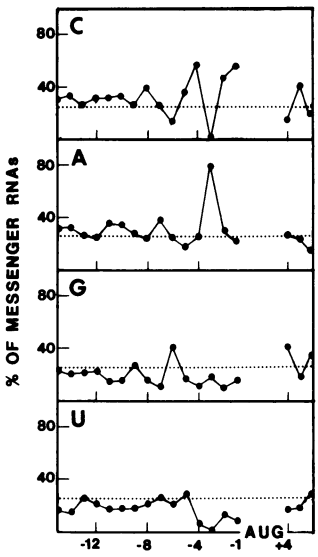

Figure 2. Frequency distribution of each nucleotide around the functional initiator codon in 198 mRNAs listed in the table. The calculations presented here do not include the 13 mRNAs in which the translational start site was predicted but has not been verified. The nucleotide immediately preceding the AUG codon is numbered -1; nucleotides +4 to +6 represent the start of the protein coding sequence. The dotted line across each panel indicates the 25% value that would be expected on a random basis. To ensure that the results were not distorted by the inclusion of too many closely-related globin mRNA sequences, I recalculated the frequency of occurrence of each nucleotide in positions -1 through -8, omitting all of the globin sequences. Although the absolute values changed somewhat (e.g. G in position -6 dropped from 40% to 36%), the rank order of nucleotides in each position remained unchanged.

the upstream AUG triplet(s) in such messages (Kozak, Microbiol. Rev., 47, 1-45, 1983; Kozak, manuscript submitted). The main point to note here is that such mRNAs are rare. The "first-AUG-rule" holds for 93% to 95% of the entries in the table.

(iii) The sequences in the table have been searched manually for signs of a conserved motif that might uniquely identify AUG initiator codons. The most conspicuous conserved feature is presence of a purine (most often A) in position -3; i.e., three nucleotides upstream from the initiator codon. As illustrated in Figure 2, 79% of the mRNAs that were counted have A in that position, 18% have G, and only 3% (a total of 6 messages) have a pyrimidine in position -3. The strong preference for a purine in position -3 is peculiar to AUG triplets that serve as initiator codons. *Pyrimidines* are favored in the -3 position preceding AUG triplets that lie *upstream* from the initiation site, in those rare mRNAs that have upstream AUGs (Kozak, Nuc. Acids Res. 9, 5233-5252, 1981); and the nucleotide frequency in position -3 is almost perfectly random around AUG triplets that code for methionine at *internal* positions in polypeptide chains (Kozak, 1983, op. cit.). Although no other position is as highly conserved as the purine in position -3, the distribution of nucleotides is decidedly nonrandom in every position from -1 through -6, and perhaps beyond. The predominance of C in positions -1, -2, -4 and -5 was evident in an earlier survey (Kozak, 1981, op cit.) and is confirmed here. The preference for G in position +4, noted in the previous survey,

is less evident here. From the data in Figure 2, the sequence CC$^{A}_{G}$CCAUG(G) emerges as a consensus sequence for eukaryotic initiation sites. The extent to which a given message matches the -1 to -5 consensus sequence varies considerably: only 10 mRNAs in the table conform perfectly to the CCACC sequence; in more than half of the mRNAs, 3 or 4 of the nucleotides directly preceding the AUG codon match the consensus sequence; about 10% of the mRNAs have a purine in position -3 but otherwise differ entirely from the -1 to -5 consensus. The 6 mRNAs in the tabulation that lack a purine three nucleotides upstream from the initiator codon do not seem to compensate by conforming closely to the other four consensus positions. Recent site-directed mutagenesis experiments (Kozak, manuscript submitted) have confirmed the importance of the purine in position -3, but there is as yet no evidence that cytosine in positions -1, -2, -4 and -5 contributes to recognition of eukaryotic initiation sites.

Obviously the (semi)conserved sequence revealed by a survey such as this need not correspond to the most favorable context for initiation, since the table includes mRNAs that vary in translational efficiency. Nonetheless, reference to the consensus sequence, *especially the highly conserved -3 position,* can be of help when searching a new mRNA sequence to locate the translational initiation site. It is important to avoid two errors when using this approach:

(a) If inspection of the sequence near the 5'-end of the mRNA were to reveal two AUG triplets that conform approximately equally to the consensus sequence, it would be incorrect to conclude that either AUG is equally likely to be the initiator codon. Because 40S ribosomal subunits most likely scan the 5'-end of the mRNA in a linear fashion (Kozak, Cell 34, 971-978, 1983), *the 5'-proximal AUG triplet is the first to be "inspected."* If the sequence preceding the first AUG triplet conforms closely to the consensus, especially if an A occurs in position -3, *the search ends there.* There are two exceptions to this rule. The first involves a small number of mRNAs in which the reading frame following the first *ANNAUG* sequence is short, terminating upstream from a second AUG codon *to which ribosomes seem to gain access by reinitiating!* The second exception consists of a single example: the mRNA derived from influenza B virus genome segment 6 allows ribosomes to initiate efficiently at the first and the second AUG codons, although the first AUG triplet occurs in a "good" context (ANNAUGA) and is not followed by a terminator codon (Shaw et al., Proc. Natl. Acad. Sci. USA 80, 4879-4883, 1983). I have no explanation for this at present.

(b) An AUG triplet that deviates from the consensus in the crucial -3 position *can nevertheless serve as the initiator codon.* This is evidenced by a few mRNAs in the table (entries 40, 98, 129, 133, 134, 196) and also by experimental manipu-

lation of the sequence flanking the initiator codon (Sherman et al., Cell 20, 215-222, 1980; Kozak, manuscript submitted). As a consequence of initiating at a "weak" AUG codon, however, those rare messenger RNAs are predicted to have two special properties: translation should be inefficient; and ribosomes should initiate not only at the first (weak) AUG *but also at the next AUG that lies downstream*. Such mRNAs should therefore have the potential to direct synthesis of two proteins. This has been shown to occur with a few viral mRNAs (Kozak, 1983, op.cit.) but it has yet to be demonstrated for cellular mRNAs.

The -1 to -5 consensus sequence detected in this survey differs from previously-suggested eukaryotic consensus sequences (Hagenbüchle et al., Cell 13, 551-563, 1978; Baralle and Brownlee, Nature 274, 84-87, 1978; Stiles et al., Cell 25, 277-284, 1981) in both its high frequency of occurrence and its constant position relative to the AUG initiator codon. None of the previously-suggested consensus sequences met those criteria. Until further experiments are carried out, it is premature to speculate about the mechanism by which flanking nucleotides might modulate recognition of the AUG initiator codon; but the temptation is irresistable. Sargan et al. (FEBS Lett., 147, 133-136, 1982) have noted an intriguing complementarity between the sequence CCACC in mRNA and the sequence GGUGG at the base of the 3'-terminal hairpin structure in 18S ribosomal RNA. The possibility of base pairing between mRNA and rRNA thus seems worth exploring. An alternative rationalization for the conserved sequence preceding the initiator codon is that it might base-pair with a complementary sequence just downstream from the AUG codon. The resulting hairpin could help to identify the initiation site. Although some mRNAs (see entries 18, 66, 151) have the potential to form a stable hairpin structure centered about the AUG codon, this is by no means universal. Moreover, comparison of closely-related sequences does not reveal compensatory changes that would preserve the potential hairpin structure.

Bibliography. The numbers used here to identify each mRNA correspond to those in column 1 of the table. Bibliographic data are presented in condensed form: first author, year, journal title, volume, first page. Personal communications are indicated by the letters *pc* after an individual's name. My thanks are extended to those individuals.

001. Noda(1982)Nature 299,793.
002. Noda(1983)Nature 302, 528.
003. Noda(1983)Nature 301, 251.
004. Ricca(1981)JBC 256, 11199.
005. Fornwald(1982)NAR 10, 3861.
006. Hanauer(1983)NAR 11, 3503.
007. Zakut(1982)Nature 298, 857.
008. Nudel(1983)NAR 11, 1759.
009. Sanchez(1983)JMB 163, 533.
010. Sanchez, op.cit.
011. Benyajati(1983)Cell 33,125.
012. Rogers(1983)JBC 258, 8169.
013. Hagenbüchle(1980)Cell 21, 179.
014. Hagenbüchle(1981)Nature 289,643.
015. Ohkubo(1983)PNAS 80, 2196.
016. Bock(1982)NAR 10,8113.
017. Amara(1982)Nature 298,240.
018. Putkey(1983)JBC 258, 11864.
019. Hennighausen(1982)EJB 126, 569.
020. Hobbs(1982)NAR 10, 8079.

021. Blackburn(1982)NAR 10, 2295.
022. Hobbs(1982)NAR 10, 8079;
Yu-Lee(1983)JBC 258, 10794.
023. Rodakis(1982)PNAS 79,3551.
024. Jones(1980)Cell 22, 855.
025. Jones, op.cit.
026. Jones, op.cit.
027. Jones, op.cit.
028. Fiddes(1980)Nature 286, 684.
029. Harris(1982)NAR 10,2177.
030. Vogeli(1981)PNAS 78, 5334.
031. Wiebauer(1982)PNAS 79,7077.
032. Jeltsch(1982)EJB 122,291.
033. Furutani(1983)Nature 301,537.
034. Snyder(1982)Cell 29,1027.
035. Snyder, op.cit.
036. Mizukami(1983)PNAS 80,3958.
037. Nunberg(1980)Cell 19,355.
R. Schimke, *pc.*
038. MacDonald(1982)Biochem. 21,1453.
039. MacDonald, op.cit.
040. Noda(1982)Nature 297,431.
Comb(1982)Nature 295,663.
041. Scott(1983)Science 221,236.
Gray(1983)Nature 303, 722.
042. Gordon(1983)JBC 258,3356.
043. Morinaga(1983)PNAS 80,4604.
044. Law(1981)Nature 291,201.
Eiferman(1981)Nature 294,713.
045. Kant(1983)PNAS 80,3953.
046. Crabtree(1982)Cell 31,159.
047. Tsujimoto(1979)Cell 18,591.
048. Yoo(1982)PNAS 79,1049.
049. Engel(1983)PNAS 80,1392.
050. Erbil(1982)Gene 20,211.
051. Erbil(1983)EMBO 2,1339.
052. Baralle(1977)Cell 12,1085.
Wilson(1980)JBC 255,2807.
053. Proudfoot(1982)Cell 31,553.
054. Baralle(1978)Nature 274,84.
055. Baralle(1977)Nature 267,279.
Heindell(1978)Cell 15,43.
056. Kay(1983)NAR 11,1537.
057. Dolan(1983)JBC 258,3983.
058. Hampe(1981)Gene 14,11.
059. Roninson(1981)PNAS 78,4782.
060. Haynes(1980)PNAS 77,7127.
061. Baralle(1977)Cell 12,1085.
062. Slightom(1980)Cell 21,627.
063. Baralle(1980)Cell 21,621.
064. Konkel(1978)Cell 15,1125.
Baralle(1978)Nature 274,84.
065. Hansen(1982)JBC 257,1048.
066. Baralle(1977)Cell 10,549.
Efstratiadis(1977)Cell 10,571.
067. Hardison(1981)JBC 256,11780.
068. Patient(1983)JBC 258,8521.
069. Banville(1983)JBC 258,7924.
070. Laperche(1983)Cell 32,453.
071. Lund(1982)PNAS 79,345.
072. Lund(1983)JBC 258,3280.
073. Bell(1983)Nature 302,716.
074. Muskavitch(1982)Cell 29,1041.
075. Garfinkel(1983)JMB 168,765.
076. Garfinkel, op,cit.
077. Garfinkel, op.cit.
078. Fiddes(1979)Nature 281,351.
Fiddes(1981)JMAG 1,3.
079. Chin(1981)PNAS 78,5329.
080. Miller(1980)JBC 255,7521.
Woychik(1982)NAR 10,7197.
081. DeNoto(1981)NAR 9,3719.
082. Page(1981)NAR 9,2087.
083. Török(1980)NAR 8,3105.
Ingolia(1980)Cell 21,669.
084-087. Ingolia(1981)NAR 9,1627.
Southgate(1983)JMB 165,35.
088. Lalanne(1983)NAR 11,1567.
Kvist(1983)EMBO 2,245.
089. Mathis(1983)Cell 32,745.
090. Malissen(1983)Science 221,750.
091. Long(1983)EMBO 2,389.
092. Sugarman(1983)JBC 258,9005.
093. D'Andrea(1981)NAR 9,3119.
094. Harvey(1983)PNAS 80,2819.
095. Grandy(1982)JBC 257,8577.
096. Sugarman(1983)JBC 258,9005.
097. Ruiz-Vazquez(1982)NAR 10,2093.
Krieg(1983)NAR 11,619.
098. Clark(1981)NAR 9,1583.
099. Heintz(1981)Cell 24,661.
100-104. Sures(1980)PNAS 77,1265.
105-108. Busslinger(1980)NAR 8,957.
Hentschel(1980)Nature 285,147.
109-110. Childs(1982)Cell 31,383.
111. Turner(1983)NAR 11,4093.
112-113. Moorman(1982)FEBS Lett.144,235.
114-115. Moorman(1981)FEBS Lett.136,45.
116. Jolly(1983)PNAS 80,477.
117. Konecki(1982)NAR 10,6763.
118. Selsing(1981)Cell 25,47.
119. Kelley(1982)Cell 29,681.
120. Bernard(1978)Cell 15,1133.
Picard(1983)PNAS 80,417.
121. Early(1980)Cell 19,981.
122. Kataoka(1982)JBC 257,277.
123. Kenten(1982)PNAS 79,6661.
124. Hellman(1982)NAR 10,6041.
125. Perler(1980)Cell 20,555.
126. Bell(1980)Nature 284,26.
127. Lomedico(1979)Cell 18,545.
Cordell(1979)Cell 18,533.
128. Hobart(1980)Science 210,1360.
129. Chan(1981)JBC 256,7595.

130. Sorokin(1982)Gene 20,367.
131. Goeddel(1980)Nature 287,411.
Lawn(1981)PNAS 78,5435.
132. Nagata(1980)Nature 287,401.
133-134. Goeddel(1981)Nature 290,20.
Lawn(1981)Science 212,1159.
135. Shaw(1983)NAR 11,555.
136. Houghton(1980)NAR 8,1913.
Ohno(1981)PNAS 78,5305.
137. Higashi(1983)JBC 258,9522.
138. Derynck(1982)NAR 10,3605.
Gray(1982)Nature 298,859.
139. Devos(1983)NAR 11,4307.
Taniguchi(1983)Nature 302,305.
140. Mason(1983)Nature 303,300.
141. Swift(1982)PNAS 79,7263.
142. Steinert(1983)Nature 302,794.
143. Powell(1983)NAR 11,5327.
144. Nawa(1983)PNAS 80,90.
145. Hall(1982)NAR 10,3503.
146. Hoffman(1982)NAR 10,7819.
147. Brisson(1982)PNAS 79,4055.
148. van het Schip(1983)NAR 11,2529.
149. Cheung(1983)NAR 11,3703.
150. McLean(1983)JBC 258,8993.
151. Grez(1981)Cell 25,743.
Jung(1980)PNAS 77,5759.
152. Vlasak(1983)EJB 135,123.
153. Karin(1982)Nature 299,797,
154. Glanville(1981)Nature 292,267.
155. Daniel(1983)EMBO 2,1061.
156. Blanchetot(1983)Nature 301,732.
157. Nabeshima(1982)NAR 10,6099.
158. Ullrich(1983)Nature 303,821.
159. VanStraaten(1983)PNAS 80,3183.
160. Watt(1983)Nature 303,725;
PNAS 80,6307.
161. Capon(1983)Nature 304,507.
162. Takeya(1983)Cell 32,881.
163. Nakanishi(1981)EJB 115,429.
164. Whitfeld(1982)DNA 1,133.
Cochet(1982)Nature 297,335.
165. McReynolds(1978)Nature 273,723.
166. Catterall(1980)J.Cell Biol.87,480.
167. Kronenberg(1979)PNAS 76,4981.
Weaver(1982)Mol.Cell.Endocrin.
28,411.
168. Hendy(1981)PNAS 78,7365.
Vasicek(1983)PNAS 80,2127.
169. Sogawa(1983)JBC 258,5306.
170. Slightom(1983)PNAS 80,1897.
171. Michelson(1983)PNAS 80,472.
172. Barrera-Saldaña(1983)JBC 258,3787.
173. Pennica(1983)Nature 301,214.
174. Cooke(1980)JBC 255,6502.
Cooke(1982)Nature 297,603.
175. Sasavage(1982)JBC 257,678.
176. Gregory(1982)NAR 10,7581.
177. Lonberg(1983)PNAS 80,3661.
178. Hudson(1981)Nature 291,127.
179. Panthier(1982)Nature 298,90.
180. MacDonald(1982)JBC 257,14582.
181. Amaldi(1982)Gene 17,311.
182. Berry-Lowe(1982)JMAG 1,483.
183. Kandala(1983)NAR 11,3169.
184. Okamoto(1982)JBC 257,15192.
185. Haché(1983)JBC 258,4556.
186. Dugaiczyk(1982)PNAS 79,71.
187. Gubler(1983)PNAS 80,4311.
188. Hobart(1980)Nature 288,137.
189. Magazin(1982)PNAS 79,5152.
190. Minth(1982)JBC 257,10372.
191. Shen(1982)PNAS 79,4575.
192. Funckes(1983)JBC 258,8781.
193. Parker(1982)NAR 10,5121.
194. Parker, op.cit.
195. Hurst(1983)EMBO 2,769.
196. Edens(1982)Gene 18,1.
197. Gurr(1983)PNAS 80,2122.
198. MacDonald(1982)JBC 257,9724.
199. Lemischka(1982)Nature 300,330.
200. Valenzuela(1981)Nature 289,650.
201. Lee(1983)Cell 33,477.
202. Suske(1983)NAR 11,2257.
Chandra(1981)DNA 1,19.
203. Itoh(1983)Nature 304,547.
204. Land(1982)Nature 295,299.
205. Schmale(1983)EMBO 2,763.
206. Geiser(1983)JBC 258,9024.
207. Hung(1981)NAR 9,6407.
208. Hung(1983)JMB 164,481.
209. Hennighausen(1982)NAR 10,2677.
A. Sippel, *pc.*
210. Pedersen(1982)Cell 29,1015.
211. Marks(1982)JBC 257,9976.