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 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.

Nucleic Acids Research

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 SI. 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
<table>
<thead>
<tr>
<th>ENTRY NO.</th>
<th>MESSENGER RNA</th>
<th>SEQUENCES FROM THE 5'-NONCODING PORTION OF EUKARYOTIC CELLULAR mRNAs</th>
<th>CAP SITE</th>
<th>AUG</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>Acetylcholine receptor pre-α-subunit (Torpedo)</td>
<td>[186]...GTTATTAGAAATGGCGAGATTATGCTGAAAGCCAAATTTTGAAAGCTGAAGAATGATTCTG</td>
<td>c b</td>
<td></td>
</tr>
<tr>
<td>002</td>
<td>Acetylcholine receptor pre-γ-subunit (Torpedo)</td>
<td>[119]...CCCTCACCACAAGCTACTCAACAGTCAGCTAGCAACTCAGCAGCCATGCTACTG</td>
<td>c b (1)</td>
<td></td>
</tr>
<tr>
<td>003</td>
<td>Acetylcholine receptor pre-δ-subunit (Torpedo)</td>
<td>[456]...GACGTGAAATCTAATGCAACGCATATAATGCGACCTGGACCTCTCTTCACATAGGGGAAc</td>
<td>c b (8)</td>
<td></td>
</tr>
<tr>
<td>004</td>
<td>α1-Acid glycoprotein (rat)</td>
<td>...CTTTCTGGCCGCTGCTGTTGCTGCTCCCGCCTGGCCTGCTCCCGGCAACCTAGACCATGCTGAC</td>
<td>c *</td>
<td></td>
</tr>
<tr>
<td>005</td>
<td>α-Actin, skeletal muscle (chicken)</td>
<td>[73]...GAGGCGCTCTCTCCAGGGGAGCGAGACAGCTACACAGGTTAGGAATATGTTGAC</td>
<td>b a</td>
<td></td>
</tr>
<tr>
<td>006</td>
<td>α-Actin, skeletal muscle (human)</td>
<td>[103]...GGCCCGAGCGAGAGTAGTGGGATGATGACCTGCTCCGGGAGGAAGACTTAGACACATGCTGAC</td>
<td>c a</td>
<td></td>
</tr>
<tr>
<td>007</td>
<td>α-Actin, skeletal muscle (rat)</td>
<td>[71]...CTTTCTGGCCGCTGCTGTTGCTGCTCCCGGAGCGAGACACTAGACACATGCTGAC</td>
<td>b a</td>
<td></td>
</tr>
<tr>
<td>008</td>
<td>β-Actin, cytoplasmic (rat)</td>
<td>[80]...ACAACCTCTCTCCAGCCCTGCTCCGGGAGCGAGACACTAGACACATGCTGAC</td>
<td>b a</td>
<td></td>
</tr>
<tr>
<td>009</td>
<td>Actin, gene 79B (Drosophila)</td>
<td>[750]...CTTGTGACTCCCTTGAACCCCTGCTACGTAGACACACACACTACACATGCTGAC</td>
<td>b a</td>
<td></td>
</tr>
<tr>
<td>010</td>
<td>Actin, gene 88F (Drosophila)</td>
<td>[188]...TGAGCTAACCTGTGACCTTCACCTCATCTCCTCAGATACACACTCAATGCTGAC</td>
<td>b a (3)</td>
<td></td>
</tr>
<tr>
<td>011</td>
<td>Alcohol dehydrogenase adult form (Drosophila)</td>
<td>[123]...TCTTAAATTGATCAAATCGGAAAGGCTGCTAAAGCAAAGAGATCGACACTAGCTGCTGTTT</td>
<td>b a</td>
<td></td>
</tr>
<tr>
<td>012</td>
<td>pre-α-Amylase (barley)</td>
<td>[96]...AAGAAAGGAGATGGTTGCTTGTGACTTGACATGAGCGGCGCCATGGGGAAG</td>
<td>c c</td>
<td></td>
</tr>
<tr>
<td>013</td>
<td>pre-α-Amylase, pancreatic (mouse)</td>
<td>GACAACTTCTACAGAAATAGTTAGTTGAGAATATCGGCAACAGCATTAGAATGAAATTC</td>
<td>a b</td>
<td></td>
</tr>
<tr>
<td>014</td>
<td>pre-α-Amylase, salivary -- liver (mouse)</td>
<td>[95]...GCACATGAAATAGTAGTTGAGAATATCGGCAACAGCATTAGAATGAAATTC</td>
<td>b a (1)</td>
<td>b a</td>
</tr>
<tr>
<td>015</td>
<td>proAngiotensinogen (rat)</td>
<td>[205]...TCAGCTGGCCCTGAGCTAAAGGACACAGCGAGAGCTCCACAGATCGCATCCCTCGGCAAGATGCTGACTCACCCTCGGACATGCTGAC</td>
<td>c b (1)</td>
<td></td>
</tr>
<tr>
<td>016</td>
<td>proAntithrombin III (human)</td>
<td>[67]...GATCAGACTTCTCTCCAGCCCGCTTGTGGAAGATTTAGGGCCGCTGATCATACCTCCCGGCAAGATGCTGACTTACCTCGGACATGCTGAC</td>
<td>b c</td>
<td></td>
</tr>
<tr>
<td>017</td>
<td>preproCalcitonin (rat)</td>
<td>[68]...CATAGGAGGACCCCGACCTCTGAGTCAATGCTCACCCAGGGAGGACATCGGCTGCTGCTGATCTCAGGGCGCCACCTAGCGAAGATGCTGACTTACCTCGGACATGCTGAC</td>
<td>c c</td>
<td></td>
</tr>
<tr>
<td>018</td>
<td>Calmodulin (chicken)</td>
<td>[91]...CGCAGCTGCAAGGACTTCAGCTCCGACGCACCCCGAGGCAGGCGCCACCTAGCGAAGATGCTGACTTACCTCGGACATGCTGAC</td>
<td>c a</td>
<td></td>
</tr>
<tr>
<td>019</td>
<td>pre ε-Casein (mouse)</td>
<td>...TAGGAAGCAGAGCTATGATGAAGATGCTGACTTACCTCGGACATGCTGACTTACCTCGGACATGCTGAC</td>
<td>c e</td>
<td></td>
</tr>
<tr>
<td>020</td>
<td>pre α-Casein (rat)</td>
<td>...GATCATCTCCACGCTTTCTCAGCTCTTGGTTCAGATCTTAGAACCACTAGAACTTT</td>
<td>c,a</td>
<td></td>
</tr>
<tr>
<td>021</td>
<td>pre β-Casein (rat)</td>
<td>ATCCCTGAGGCTTCATCTTCTCTCTCTGCAGTGTACAGCATGGCATGATGGTTG</td>
<td>b,a</td>
<td></td>
</tr>
<tr>
<td>022</td>
<td>pre γ-Casein (rat)</td>
<td>ATCATCTCTACCTATTCCTCTCTCTGTCCTCCGCTAAAGGACTTGACAGCCATGAAGGTC</td>
<td>b,a</td>
<td></td>
</tr>
<tr>
<td>023</td>
<td>prechorion proteins (silkworm) - m1911</td>
<td>...CTGAATATTCCACGACATCATGTTTACA</td>
<td>c,a,c</td>
<td></td>
</tr>
<tr>
<td>024</td>
<td>-gene 292a, A-family</td>
<td>ATCATTCTAGTTCAACGGAGTCTCATGCTAATGACAGATGGCATG</td>
<td>b,b</td>
<td></td>
</tr>
<tr>
<td>025</td>
<td>-gene 18b, A-family</td>
<td>GTCATTCTGAAATTATATCTCATTAGCAGAACATGGTACCA</td>
<td>b,b</td>
<td></td>
</tr>
<tr>
<td>026</td>
<td>-gene 40a, B-family</td>
<td>ATCATCTCAGTTGATTTCAAAATGAAACATGACTCAC</td>
<td>b,b</td>
<td></td>
</tr>
<tr>
<td>027</td>
<td>-gene 10a, B-family</td>
<td>ATCATATTGGATTTCTCTCTCAACAAAGAAATGCGAC</td>
<td>b,c</td>
<td></td>
</tr>
<tr>
<td>028</td>
<td>Chorionic gonadotropin (&gt;127)</td>
<td>...AGCAAGCCGAGGGGACGCACCAAGGATGGAGATG</td>
<td>c,b,c</td>
<td></td>
</tr>
<tr>
<td>029</td>
<td>preprochymosin (bovine)</td>
<td>AGCAGCCGCTGAAACTCTGAGGATGACATGGGTTCGA</td>
<td>b,c</td>
<td></td>
</tr>
<tr>
<td>030</td>
<td>preprocollagen-α2, type I (chicken)</td>
<td>...CCACGTCCGGGGGCTCTGCAACACAAGGAGTCTCATGCTAATGACAGATGCTAC</td>
<td>b,c</td>
<td></td>
</tr>
<tr>
<td>031</td>
<td>precomplement, C3 (mouse)</td>
<td>ATCATTCTGACATTCTACCTCTCTGTCCTCCGCTAAAGGACTTGACAGCCATGAAGGTC</td>
<td>b,a</td>
<td></td>
</tr>
<tr>
<td>032</td>
<td>precomplement (chicken)</td>
<td>ATCATTCTGACATTCTACCTCTCTGTCCTCCGCTAAAGGACTTGACAGCCATGAAGGTC</td>
<td>b,a</td>
<td></td>
</tr>
<tr>
<td>033</td>
<td>preproctoecitropin releasing factor (ovine) (&gt;127)</td>
<td>...AGCAAGCCGCTGACTCTGACTCTGACAGGGATGGAGCTAC</td>
<td>c,d</td>
<td></td>
</tr>
<tr>
<td>034</td>
<td>precuticle protein I (Drosophila) (&gt;127)</td>
<td>ATCAGTCCAAGGTTCCTGCAACAGGAGTCTCATGCTAATGACAGATGCTAC</td>
<td>b,b</td>
<td></td>
</tr>
<tr>
<td>035</td>
<td>precuticle protein III</td>
<td>ATCAGTCTAGAAGATTATCTCAGTGACAGGACACCTACCCAAATCAGATCGGTCA</td>
<td>b,b</td>
<td></td>
</tr>
<tr>
<td>036</td>
<td>cytochrome P-450 (rat)</td>
<td>ACTGAAGTCTCATGCTCAGGACACCTACCCAAATCAGATCGGTCA</td>
<td>b,b</td>
<td></td>
</tr>
<tr>
<td>037</td>
<td>dihydrofolate reductase (mouse) (+ longer forms) (&gt;127)</td>
<td>ATCAGTCTAGAAGATTATCTCAGTGACAGGACACCTACCCAAATCAGATCGGTCA</td>
<td>b,b</td>
<td></td>
</tr>
<tr>
<td>038</td>
<td>prepro elastase I (rat)</td>
<td>GTCAGTCTACTCTCAGTCCAAACACTGCTGAC</td>
<td>b,c,b</td>
<td></td>
</tr>
<tr>
<td>039</td>
<td>prepro elastase II (rat)</td>
<td>ACAGACATCCAGGAACACACATCCAGTCAGG</td>
<td>b,c,c</td>
<td></td>
</tr>
<tr>
<td>040</td>
<td>prepro enkephalin (human) (&gt;127)</td>
<td>CTGAACGGCCTTTTCCAATTGGCCTGCTCATCCGAACAGCGTCAAC/TCCATGGCGCGG</td>
<td>b,c</td>
<td></td>
</tr>
<tr>
<td>041</td>
<td>prepro epidemal growth factor (mouse) (&gt;127)</td>
<td>TCAGAGGCTCTGAGGGTGAAGCATCTGAGCAGGACATCCAAAGGAGGAGCAGTAAATAGATGCGCTGG</td>
<td>b,c,d</td>
<td></td>
</tr>
<tr>
<td>042</td>
<td>fatty acid binding protein (rat liver) (&gt;127)</td>
<td>CTGTGGTGGCAGCTGGGAAAGGAAACCTCATTCGAGCACCCTAGACTCTC</td>
<td>c,a</td>
<td></td>
</tr>
<tr>
<td>043</td>
<td>pre α-fetoprotein (human) (&gt;127)</td>
<td>ATTTGTGCTTCCACCACTCCACTCGAATACAAATATACGCAACATGCTGAC</td>
<td>b,c,c</td>
<td></td>
</tr>
<tr>
<td>044</td>
<td>pre α-fetoprotein (mouse) (&gt;127)</td>
<td>ACATCCCACTCCAGCTCAGGCGTGAAGGAGACAGCAGGACATGCGAGG</td>
<td>b,c,a</td>
<td></td>
</tr>
<tr>
<td>045</td>
<td>pre fibrinogen, αa (human) (&gt;127)</td>
<td>TCTCTTTTCTTGCAGCTGAGTGCTCAGCAGCAGCTGGCCAGGCCCCACCTACTTCTAGAAAGATGTTTCC</td>
<td>c,a</td>
<td></td>
</tr>
</tbody>
</table>
pre-Fibrinogen, γ (rat) [53] AGAGGTCAAGCAGGCTGGCTGATAGGGGCTGATACTGCACTCAGACACACTAGATGATTG b,e b
preproFibrin (silkmoth) [24] ACATCTTGCTGCCAATCTCAGAGTGAGGCAGTGCAGACACTAGATGATTG a,b c
preproGastrin (porcine) [61] ACTGGGACCAGGCAACAGCAGCAGCAGTGGCTGATCGGAGCACACTAGATGATGA c c (l)

α-Globin family:
- chicken, embryonic (W') [55] ACAACCTGCTGTGTTGTTCTCAGTGAAGGGGCAAGGCAACACTCCTCTGACACATGACGACTG d,e a
- duck, major adult (αA) [36] ACCCGTCTGGGCTGACAAGGAGCTGACAACATGATGATTG b a
- duck, minor adult (αD) [42] ACAGAAGCCAGTCACTAGCCACACGCGGCTGGCAGATGACGACTG b a
- human, adult [37] ACCTCTGTGGACCAAGATGACGACTG b a
- human, embryonic (λ) [55] ACAAGGCGGTCGCTGACAGGCAACACTCCTCGACAGGCACACTAGATGATTG a a
- mouse, adult [32] ACTTCTGTTCTGAGACTGACAGAGAACACATGACGACTG a a
- rabbit, adult [36] ACACCTTGCTGACAGGACAGAAGCAGCAGCAGCAGATGACGACTG a a
- Xenopus, major adult ...TGCCAAAGAACAACGAAAACACTAGATGATTG c c

B-Globin family:
- chicken, adult [77] ...GAGCCCGAGCCCCTCGTATCCGAAAACAGCAGGCTACCTCCTAACCCTCCGGAGGATGACGACTG b a
- duck, adult [45] ...AGCCGAGACCTCCCGTATCCGAGGACAGGACACTCCTCCTAGGCGGCAAGGATGACGACTG c a
- chicken, embryonic (p) [45] AGAAGCTGCTGCGCTGTGCTGACCCCCGACAGGATGACGACTG c a
- goat, adult (βB) [52] AACTTCTGGCTGACAGGATGACGACTG b a
- human, adult [50] AACTTCTGGCTGACAGGATGACGACTG a a
- human, fetal (γ) [53] AACTTCTGGCTGACAGGATGACGACTG a a
- human, embryonic (ε) [53] AACTTCTGGCTGACAGGATGACGACTG a a
- mouse, major adult [52] ACTTTTGTGGCTGAGACTGACAGGATGACGACTG a a
- mouse, embryonic [52] ACTTTTGTGGCTGAGACTGACAGGATGACGACTG a a
- rabbit, adult [62] ...TCTGAGACATGACAGGATGACGACTGAGGAGATGACGACTG a a
- rabbit, embryonic (E3) [62] ...TCTGAGACATGACAGGATGACGACTGAGGAGATGACGACTG a a
- Xenopus, major adult [46] ACTTGGTCTCCTCGCAGGATGACGACTGAGGAGATGACGACTG a a
- Xenopus, larval ...TAGAAGCTGCTGGGGATGACGACTGAGGAGATGACGACTG a a

pre α2Globulin (rat) [68] ...CCATGAGAGAGGATGTTGCGGACAGGAATGGTTCTTCTCAGACAGGATGACGACTG a a
preproGlucagon I (anglerfish) [58] ...AGGAACTAAACAGCAGCATTGGGGGTGTTG a a
preproGlucagon II (" ) ...GAAGCTTCAAGAATGACGACTG a a
preproGlucagon (hamster) [103] ...TGCCAGCAGGTCACCTCCTCGCGCTGCAGGAGGAGGAGATGACGACTG a a
preGlue proteins (Drosophila)
- Sgs-4 [13] TTCCAAAGGTCAAGGATGCGCTG b d
- Sgs-3 [29] ATCGCTTTTGGAGGAGTTAATGAAAAACAGGATGACGACTG c,d a
- Sgs-7 [33] ATCTTGAAAGATGTCTTCTAGAAGGAGGAGATGACGACTG b,d a
- Sgs-8 [33] ATCTTGAAAGATGTCTTCTAGAAGGAGGAGATGACGACTG b,d a
Glycoprotein hormones (human) [100] ...CATCAAGGAGGAGGTCCACAGATGACGACTGAGGAGATGACGACTG a,d b
pre-α-subunit
Glycoprotein hormones (mouse):  
- pre-a-subunit
- pre-Growth hormone (bovine)
- pre-Growth hormone (human)
- pre-Growth hormone (rat)

Heat shock (Drosophila):  
- 70K protein
- 22K protein
- 23K protein
- 26K protein
- 27K protein

Histocompatibility (MHC) antigens:  
- Class I: mouse (pre)H-2Kd
- Class II: mouse (pre)I-Ag
- Human (pre)DR

Histones:  
- Chicken H1, embryonic
- H2A.1, adult
- H2A.F, embryonic
- H4, embryonic
- H5, embryonic

Histones:  
- Human H3
- H4

- Sea urchin, early H1 (S. purpuratus)
- H2A
- H2B
- H3
- H4

- Sea urchin, early H2A (P. miliaris, h22)
- H2B
- H3
- H4

- Sea urchin, late H3 (L. pictus)
- H4

- Xenopus H1
- H2A
- H2B
- H3

- C. elegans H1

Nucleic Acids Research
116 Hypoxanthine phosphoribosyltransferase (HPRT), human

117 HPRT, mouse

118 Immunoglobulins:

119 - (pre)γ light chain, mouse [29]

120 - (pre) light chain, mouse [3; 18]

121 - (pre) heavy chain, mouse [45]

122 - (pre) heavy chain, human IgE

123 - (pre) heavy chain, rat IgE

124 preproinsulin (chicken)

125 preproinsulin (human)

126 preproinsulin (rat-I)

127 preproinsulin (anglerfish)

128 preproinsulin (hagfish)

129 preproinsulin (salmon)

130 Interferons:

131 - pre-a (LeIF-A), human [67]

132 - pre-a (LeIF-D), human [67]

133 - pre-a (LeIF-C), human [67]

134 - pre-a (LeIF-H), human [69]

135 - pre-a (leukocyte), mouse [67]

136 - pre-β (fibroblast), human [75]

137 - pre-β (fibroblast), mouse

138 - pre-γ (immune, human [128]

139 preinterleukin-2, human [47]

140 preproallikrein, mouse [42]

141 preproallikrein, rat

142 Keratin, epidermal (mouse)

143 Keratin, B2A (sheep)

144 preKininogen, low MW (bovine) [133]

145 - pre-α-Lactalbumin, guinea pig [84]

146 prelectin (French bean) [16]

147 Leghemoglobin (soybean) [49]
| 148 | apolipoprotein II, VLD, chick [77] | GAAAGCAGGACAG/GTCTTTGCTAAAGGCTGAACTGTAACCACAACCACTGCTGCAA | b a |
| 149 | apolipoprotein A-I (human) | ...TCCCCACGCGCCCTTCCAGGATGAAAGCT | c a |
| 150 | apolipoprotein E (rat) | ...ACTGGCAATCCAACTGGAAGAGATGGAAGCT | c a |
| 151 | prelysozyme (chicken) [29] | AGTCCCGCTGTGTCGACGACTGGCAACATGAGGTCTT | b a |
| 152 | preproMelittin (honeybee) | ...AGCGAAATTAAACAGAATTACAGGAAAGGAAGAGGGAAGGCTGAGGAAATATGCTAAATTC | b a |
| 153 | Metallothionein-II (human) [69] | ...CCAGCGAACCCGCGTCAGGAACTGCTCGCGCTCTCATGCGCAATGATCCC | b a |
| 154 | Metallothionein-I (mouse) [73] | ...CTGAGTACCTTCTCTACACTACCTGCTAGCTCCAGTCTACACAGATCTGGAAATGCC | b a |
| 155 | pre-82-Microglobulin (mouse) [52] | ...ATTTACGCTGTGCTACTGCGCGTCTCCGCGCTCTACTTTTCCAACTCTCAATCATGTCCTTC | c d |
| 156 | Myoglobin (seal) [70] | ...CAGGGCAAGGCTGACCGGACTGCTCTTCTCTCTTCTCCAGACTGACCATGAGGGCTC | b a |
| 157 | Myosin, skel. L chain (chick) | ...CTTCAGCTCAATCCCTGGGCGCGCTCTACTTTTCTACAATCTCAATCATGTCCTTC | c c |
| 158 | prepro-8-Nerve growth factor [74] | ...GCTGGCCTTATATTTGATCTGGGATTTTTTTCGGGTAGTGGAAAACCAG/CAGCCTCCCGGGTGGAGTCCC | b a |
| 159 | -c-fos (human) [154] | ...CCCACCTGTCTCAGGGCCCTCCGCCTCCGCGCCCTGCCTGTAACCAACCCAGATGATGTC | e |
| 160 | -c-myc (human) [572] | ...AGACGCTGTAATTCTTTGGATGAGTGGAAACTGCTAGTGAACATGCTGTGC | c d |
| 161 | -c-Ki-ras2 (human) [181] | ...GGGCGCAGGGCTGACCGGACTGCTCGCGCTCTCATGCGCAATGATCCC | b c |
| 162 | -c-src (chicken) [128] | ...TCCCAGCAGCTGCTGCTCTCTCGCTCTACACACACACACAGAGTGGCAATGGCTGCA | b a |
| 163 | preproOpiomelanocortin (bovine) | ...GAGGGCCGCGGACGCGCTCTCCCGGTGACAG/AGGCTAGGCTCGCCGAGATGAGGAGA | a |
| 164 | preproOpiomelanocortin (human) [107] | ...CCCAGGCCCTCAGGAGGAGGAGCAGAGAGAGAGCCAGCTGATGCTGCCTGCCTGCCTGCCTGCCTGCCTGCCTGGAAGATGCGGAGA | d e a |
| 165 | Ovalbumin (chicken) [64] | ...AAACTGATTTGGCTTTTAGAGTGGTCAAGCTGAAAG/ACAACCTGAGCTACCACATGGGCTCCT | a b a f |
| 166 | preOvomucoid (chicken) [53] | ATCTCGAGGAGCAAGGGCCGAGCCTCGCAGGGGAGTACCTACCTCACATGCGGATAGGG | b a |
| 167 | preproParathyroid hormone (bovine) [100] | ...TAGGCTCTAATTGAGGAGATTTTGGTAAAGTGTCTATGCGTCTG/TTAATATGCTGCT | b |
| 168 | preproParathyroid hormone (human) [90] | ...AGTCAACTAAACTACCTGAAAGGAGATCTGCTTAAACATGTTATG/TGAAAGATGATACCT | d e (1) |
| 169 | prePepsinogen (human) [55] | ...CTGGACACCCCTCCGCTTGGCTCTTCCTCCTGACCTGGAGCTTGGGACCAGGAGAACATGGAGTGG | b e |
| 170 | pre(?)Phaseolin | [77] | ...CTTCCACCTGCTAGCCGACCTCAGCTCAGCTCAGCAGCTTCCGCTGAAATTC | a |
| 171 | Phosphoglycerate kinase, human [80] | ...GGCTCCCTCGTGGTGGAGGCTCCCTCCGCTGAAATTC | c a |
| 172 | prePlacental lactogen, human | ...CTGGGACAGGCTACCTGGCGAATGCGTCA | c a |
173 pre-Plasminogen activator (human) [84]...GGCGAGAAGGGAAGGGAAGGCAAGCCCTGTAATTTAAGGAGCCGCTGTGAAGCAATCATCATGTGAGATC

174 pre-Prolactin (rat) [51] AGTGGTTCTCTTGGGAGAAGTGTGGTCCCAGTGGTCATCACATGGACACAGGAG

175 pre-Prolactin (bovine) [67]...GGCCATAGGACGAGAGCTTCCTGGTGAACTCAACATGGACAGGAGGAGGAGTCCCTG

176 Protamine (trout) [14] ATCCCATCAATGCACTAGGCCAGA

177 Pyruvate kinase (chicken) [80]...GGCTTTGGCAACGGCGCGGCGGAGAGTTGCAGAAGCTACATGGAGACG

178 preproRelaxin (rat) [60]...CTGAGGACCTTTTCTGGCAAGATGAAGTCT

179 preproRelaxin (mouse) [58]...CAGAGAGTACGCTTGAGAAGCTACATGGAGACG

180 preproRibonuclease, pancreas (rat) [97]...CAATTCTGGCCAGGGAAGTTGCAGAAGCTACATGGAGACG

181 Ribosomal protein S19, Xenopus [46]...AATCTTTTAACTTTTTCTGGCAAGATGAAGTCT

182 preRibulose bisphosphate carboxylase (soybean) [45]...ATGCAAGCAGTCTTGAGAAGCTACATGGAGACG

183 Seminal vesicle pre-secretory protein IV (rat) [22]...ATATCTTTTCTGGCAAGATGAAGTCT

184 pre(?)-Serin (silkworm) [54]...ATAGTCCCTTCTATACGCCGCTAGTCAAGATCGCCACCAATCTGGACTCTCC

185 Serum preproalbumin, chicken [41]...AGCACCTTTTCTGGCAAGATGAAGTCT

186 Serum preproalbumin, human [41]...AGCACCTTTTCTGGCAAGATGAAGTCT

187 preproSomatocrin (human) [112]...TGCCCAGGCAAGAGTACGCTTGAGAAGCTACATGGAGACG

188 preproSomatostatin-II (angelfish) [59]...CAGAGAGTACGCTTGAGAAGCTACATGGAGACG

189 preproSomatostatin-22 (catfish) [59]...CAGAGAGTACGCTTGAGAAGCTACATGGAGACG

190 preproSomatostatin-14 (catfish) [114]...GCCCCTCCCTCCGCAAACTTTTTCCGCAAACTTTTTATTCTGGCAAGATGAAGCTACATGGAGACG

191 preproSomatostatin-I (human) [105]...GCCCCTCCCTCCGCAAACTTTTTCCGCAAACTTTTTATTCTGGCAAGATGAAGCTACATGGAGACG

192 preproSomatostatin (rat) [81]...CTGAGAGTACGCTTGAGAAGCTACATGGAGACG

193 pre-StEROID binding protein (Cl, rat prostate) [44]...CTGAGAGTACGCTTGAGAAGCTACATGGAGACG

194 pre-StEROID binding protein (C2, rat prostate) [42]...CTGAGAGTACGCTTGAGAAGCTACATGGAGACG

195 pre-StEROID binding protein (Cl, rat prostate) [55]...CTGAGAGTACGCTTGAGAAGCTACATGGAGACG

196 preproThymosin (rat) [31]...CTGAGAGTACGCTTGAGAAGCTACATGGAGACG

197 Thyrotropin, pre-β-subunit (rat) [89]...CTGAGAGTACGCTTGAGAAGCTACATGGAGACG

198 preproOxytocin (rat) [12]...CTGAGAGTACGCTTGAGAAGCTACATGGAGACG
<table>
<thead>
<tr>
<th>Entry</th>
<th>Protein Name</th>
<th>mRNA Length</th>
<th>5'-Noncoding Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>199</td>
<td>α-Tubulin (rat)</td>
<td>100 nucleotides</td>
<td>AACACCTCCTCCTCGCCACTCCACCCGGAGCGGGAGCAGCAACACCATG/CGTGAG</td>
</tr>
<tr>
<td>200</td>
<td>β-Tubulin (chicken)</td>
<td>87 nucleotides</td>
<td>AGAGCGGGAGGTGACGGAGCGGGAGCAGCGCGGCACCGGCAGACACCCGGCATGTGAG</td>
</tr>
<tr>
<td>201</td>
<td>α-Tubulin (human)</td>
<td>159 nucleotides</td>
<td>TTTTCTTGCACCCCATAACATCTCGAGGCGGACAAAAAAATATTATTTAAACCATGAGGAA</td>
</tr>
<tr>
<td>202</td>
<td>preUteroglobin (rabbit)</td>
<td>47 nucleotides</td>
<td>AGATCCAGCGATCCAGAGCCAGCCCGAGCTCCATCGGCCATGAACTC</td>
</tr>
<tr>
<td>203</td>
<td>preproVasoactive intestinal polypeptide (human)</td>
<td>108 nucleotides</td>
<td>GGGAGCAGCAGCAGAGCGGAGAGCGAGCACGACCTATGTCGCCCCATGAGGAC</td>
</tr>
<tr>
<td>204</td>
<td>preproVasopressin neurophysin (bovine)</td>
<td>49 nucleotides</td>
<td>GGCAAGTCTACAGGACAGCAGACTGCGAAGTGCTGGCCGCCAGGATGCCGAC</td>
</tr>
<tr>
<td>205</td>
<td>preproVasopressin neurophysin (rat)</td>
<td>104 nucleotides</td>
<td>AGCAGACAGAGCTGAGCTGAGCCATGCCATGAGGAC</td>
</tr>
<tr>
<td>206</td>
<td>Vitellogenin II (chick)</td>
<td>13 nucleotides</td>
<td>ATTCACCTCTGCTATGAGGGGG</td>
</tr>
<tr>
<td>207</td>
<td>Vitellogenin I (Drosophila)</td>
<td>258 nucleotides</td>
<td>ACTCACACTGCGAGCTGCGATCCCGAGACCAATCTCAACCATGAGCAGCC</td>
</tr>
<tr>
<td>208</td>
<td>Vitellogenin II (Drosophila)</td>
<td>251 nucleotides</td>
<td>ATGCAGTACAATTTGGTACGGTGTTCAAGTTCACACTGAGGAGCAACGATGATC</td>
</tr>
<tr>
<td>209</td>
<td>preWhey acidic protein (mouse)</td>
<td>26 nucleotides</td>
<td>ATGCAGTACAATTTGGTACGGTGTTCAAGTTCACACTGAGGAGCAACGATGATC</td>
</tr>
<tr>
<td>210</td>
<td>preZein, 19K (maize)</td>
<td>57 nucleotides</td>
<td>CACATATTATTGGAGACCAACTGAGCTGAGACTGAGGCAGCTAGCAAC</td>
</tr>
<tr>
<td>211</td>
<td>preZein, 22K (maize)</td>
<td>67 nucleotides</td>
<td>TCAGCATTCCAAAAACACACCAGGGAAGCTGAGGACAGCAGTACACACCTGAGGCTACTC</td>
</tr>
</tbody>
</table>

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 5'G 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.
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 $\text{CC}_{\text{G}}^{\text{A}}\text{CCAUG(G)}$ 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 table 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 (ANNAUG) 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 irresistible. 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.