

have a dominating role; nucleotide replacements elsewhere in the consensus do affect initiation only in the absence of a purine at -3 (6).

In cases where two AUG codons are found, or introduced by means of mutagenesis (5,9), in the 5' proximal region, the AUG used most frequently for protein initiation is the one placed in the most optimal context, i.e. the sequence most resembling the consensus. An upstream located AUG in a poor context has less effect on initiation of a downstream AUG in an optimal context than an upstream located AUG in optimal context (5,9).

In the compilation of Kozak (1,2), lower eukaryotes were excluded. In this paper we compare 99 translation initiation sites of the lower eukaryote Saccharomyces cerevisiae. This comparison reveals that initiation sites in this yeast are distinctly different from those in mammalian cells and also from those in plant cells. The entire untranslated sequence preceding the startcodon is very rich in A-residues. T- and C-residues appear at close to normal frequency, but G-residues are rare.

RESULTS

In the following compilation only sequences of Saccharomyces cerevisiae are included that were present in Genbank dated August 4, 1986. Thus, 96 sequences could be compared. A window of 100 bases around the startcodon was analyzed (Table Ia and Ib). Note that the transcribed sequences upstream of the AUG start are often longer than the 47 base "window" and have been truncated to make the table more readable. It should also be noted that several yeast genes are now known to have multiple transcription start sites (10-13); we have used the sequences for the upstream transcribed regions as found in the Genbank database. For the genes SUC2 and SUC7 tandem AUG triplets are found at the start position and the second AUG has been aligned as the actual start site as referenced in the Genbank database. For some ribosomal protein genes and the actin gene, the startcodon is on a different exon than the body of the coding sequence. In such cases the separating intron was omitted and the actual translated mRNA sequence is used. All sequences use AUG as startcodon; no exceptions were encountered. Table Ia shows that the sequence immediately preceding the startcodon is extremely rich in A-residues, at most positions an A-residue is present in more than 40 percent of all cases. This is in contrast to mammalian initiation sequences where A prevails only at -3, whereas C-residues are abundant at -1 and -2 and to a lesser extent also at -4 and -5. In yeast, the highest occurrence of A-residues is, as in mammalian mRNAs, at position -3 (in 81 percent of all cases).

Table 1^a Nucleotide sequences preceding the start codon

ACT	GCUUACUGCUUUUUUCUCCCAAGAU CGAAAUUU	A	CUGAAUUUA	A	C	A	A	U	G
ADE4	AUAAGUUUAGCAAAGAAAGAGGUACAGCAAAACAGC	A	G	A	A	U	A	G	A
ADH1	CAAUUAUUCAAGCUAUUACCAAGCAUACAUAUCAACU	A	U	C	U	C	A	U	A
ADH3	GUUCACAGUUA AAAACUAGGAAUAGCAUUAAGUCA	A	G	A	U	U	A	A	G
ADR2	AAAAGCAUACAAUACCAUCUAUCAUAUAUCAUAUA	A	U	A	U	A	U	A	G
ARG4	UUGAAGAGCUCAAAAGCAGGUAUCUAUUAACA	A	G	A	C	U	A	U	G
CCS	AAAAGAAAUAAGGCAAAACAUAUAUCAUAUAUAUA	A	U	A	U	A	U	A	G
CDC28	AGCUACAGUGGAAAUAAGCCAGAUCAAAUAGA	A	C	U	A	U	C	U	G
CDC8	UCUAAAACUAAUAUGAUUCAUAGUGGACAGAAAGA	A	C	A	U	A	C	A	G
CPA1	UUAAUAUCAGAAAUUUCUAUCACAAACCACUCCUA	A	A	A	A	U	U	U	G
CPA2	-----AGGAAGAGCAUAACAUAUAUA	A	G	A	G	A	G	A	A
CPAx	UUAAUAUCAGAAAUUUCUAUCACAAACCACUCCUA	A	A	A	U	U	U	U	G
CPB1	UUUAGGAUAGCAGUAGUUUGCAUUUUGCACGUUUC	C	U	U	C	C	A	U	G
CUP1	AUUAGAUAACAAAACUGUAACAUAUCAUAUAUA	A	C	A	U	C	A	U	G
CYC	UCAUCACUAUUUUUUUCAGAAAGCGGAAGUUAUA	A	A	C	U	A	U	U	G
CYC1	UAUUCUUCUAUJAGCAGCAACACACAUAUAUCA	A	C	A	C	A	U	A	G
CYC17	AUUAACACUUAACCCUACACACACGGAUAAAGAA	A	A	C	A	U	A	U	G
CYC1x	UAUACUUUCUAJAGCAGCAACACACAUAUAUCA	A	C	A	C	A	U	A	G
CYC4	CACCCAUUUCGAUUUUGAUGUUGCCAUACAUAUA	A	G	A	U	A	A	C	A
CYC7	UUACUUUAAGUAACUUCAGUUAACAUAUAUAUA	A	C	A	U	A	A	C	A
CYCr	UUUUCAGAAUUAUJAGUAACUGUAUAUAACA	C	G	U	U	C	A	A	G
EF1a	UACUUGUUUUUAGAUAUAUCGGUCAACGAUAUA	A	U	A	U	A	U	A	G
EF1ab	-----AAGCAUAGCAUUCUAUAUAUAUAUA	A	U	U	U	A	U	A	G
ENOa	---AAAACCAAGCAACUGCUUAUCAACACACA	A	A	C	A	C	A	U	A
ENOb	-----AACCCAAGCAUAUAUCAUAUAUAUAUA	A	C	A	U	A	U	A	G
G3PDa	---ACACCAAGAACUUAGUUUGAUAUAACACAC	A	C	A	U	A	A	A	G
G3PDb	AGAACUUJAGUUUCAAAUUUAUUUCAUCACACA	A	A	C	A	A	C	A	A
G3PDc	AACACACAGAAAAACAGUACUUCACUAUAUUU	A	C	A	C	A	C	A	A
GAL	AUUGUUAAUAUAACCCUCUAUAUCUUUAAC	G	U	C	A	A	A	C	A
GAL10	-----	-----	-----	-----	-----	-----	-----	-----	-----
GAL1p	-----	-----	-----	-----	-----	-----	-----	-----	-----
GAL4	-----	-----	-----	-----	-----	-----	-----	-----	-----
GAL80g	GCGUAUAACAUCUCGUAUGUUGGUUUCCGUUCU	U	C	C	G	U	C	A	G
GCN4	UGUUUACAAUUUGUCUGUCUAGAAGAAUAUAUA	A	U	A	U	A	A	U	G
GDHm	AUCGCAUUAUUCUAUAUAACAAGUUAJGGAGAC	A	A	A	U	A	A	A	G
H2A1	AAUAAACAACUUCAAAACAACAUAUAUAUAUA	A	U	A	U	A	U	A	G
H2B1	AUCGUAUAUJAGCAAGUCAAAACAACAUAUAUA	A	U	A	U	A	A	A	G
H2B2	UUCGUAUUGCUCUAUCUCAAACAACAUAUAUA	A	U	A	U	A	U	A	G
H3cI	UCUAUUAUAUJAGAAAAACAUCUAACAUAUAUA	A	U	A	A	C	G	C	A
H3cII	CAACUGUUUCUUCUUUUUAUAJAGUCCAUAUA	A	A	C	U	C	A	A	G
H4cI	ACUAAAGCAACAACAACAACAACAACAUAUAUA	A	U	A	U	A	U	A	G
H4cII	AUCUUUGUCAAAGAGUAGCAAAAACAACAUAUA	A	U	C	A	U	A	A	G
HIS1	-----AGAAAAGAAJAGUAGGUUCUAUAUAUA	A	U	U	C	A	U	A	G
HIS3	-----AGCAGGCAAGUAUAACGAJGGCAAGAU	A	A	C	A	G	A	A	G
HIS4	GCUGUUAUAUJAGUAUAUAUAUAUAUAUAUA	A	U	U	U	U	U	U	G
HMLa11	-----UCAACCAUUAUAUAUAUAUAUAUAUA	A	G	C	A	U	U	A	G
HMLa12	AUCAGCUUJAGUAGGUCAAGAAAAAAGGAJAG	A	A	G	C	A	A	A	G
HSP90	AGAAAAUJAGUUCUAUAUAACAAGCAAAACA	A	A	C	A	C	G	C	A
HXK1	---CAAACUCACCCAAACAUCUAUAUAUAUAUA	A	C	U	G	A	A	A	G
HXK2	UJAGAAUUAUUAUUCACCAUAUAUAUAUAUA	A	U	A	U	A	U	A	G
LEU1	GACAGUUUUUUGCUCUAUCUAUAUAUAUAUA	A	U	U	U	U	U	U	G
LEU2	UUACAUUUCAGCAUAUAUAUAUAUAUAUAUA	A	U	U	C	A	G	U	A
MATa11	-----UCAACCAUUAUAUAUAUAUAUAUAUA	A	C	G	A	U	U	C	A
MATa12	-----AAGAAAAAAAGGAJAG	A	A	G	C	A	A	A	G
MES1	-----CGGAUAUUUJACCAACA	A	A	U	U	C	A	A	G
MEF	AAGAUUAACAACUAUCAUUUUAUAACAUAUAUA	A	A	C	A	U	A	A	G
MEF1g	AAGAUUAACAACUAUCAUUUUAUAACAUAUAUA	A	A	C	A	U	A	A	G
MEF2g	CUACCAUCAACUGCAUCAUUUUCAGUAUAUAUA	A	C	A	U	U	U	A	G
ODCd	UAACCCAAACUGCACAGAACAAAACCCUGCAGGA	A	C	G	A	A	G	A	G
ODCf	UAACCCAACUGCACAGAACAAAACCCUGCAGGA	A	C	G	A	A	A	A	G
PGK	AUCAUCAAGGAAJUAUAUAUAUAUAUAUAUA	A	C	A	U	U	A	A	G
PHO3	UCAUAAAAAAAUJACUAGUAAGAAAGGGCC	A	U	U	C	A	U	U	G
PHO5	GJAGAAUAJAGAACCAACAUAUAJAGCAJGCA	A	A	U	U	C	G	A	G
POR	-----AAGCGUACCCAAAGCAAAAUAUAUAUA	A	C	A	C	C	A	C	A
PPR1	AGUAAACCUUJAGUACUAUAACAUAJCGAJAG	A	U	G	A	U	A	U	A
PPR2	AUCCACCUCAUAUAUUUGCAUAUCGCAAJGGC	A	U	A	U	G	A	A	G
PUT2	AUAUAUAUAUAUAUAUAUAUAUAUAUAUAUA	A	U	A	U	A	A	A	G
PYK	-----GACACCAUAUAUAUAUAUAUAUAUA	A	A	A	C	A	U	C	A
RAD1	AUGUGUAAAAUAUAUAUCUAUCUUGUA	A	A	A	U	A	U	C	A
RAD2g	-----GGGAGAAJCCCAAUUCUUA	A	C	A	U	C	A	A	G
RAD3	AGGUGAUUAUCJGGCCGAUUUGCAUACUUUA	A	G	A	A	A	U	G	A
RAD6	AAUUCAAAGAUUAUUUUAJGGCAAGACAGACUA	A	A	C	A	U	A	A	G
RASH1r	AUUUCAAGUAJAGCAGGUAAACAUAUUUUC	A	U	U	U	U	A	A	G
RASH2r	ACCAAGUUAAACCGUUAUGAAUUAJAGGAAU	A	U	A	C	A	A	A	G
RP13	UCGGUUUUGUCAUCUUCUJAGAACACAACAGU	A	C	A	C	A	U	C	A
RP29	-GCCAAGUUAJAGGGAAGACCAAGAACAUAUA	A	C	U	G	A	G	A	G

Table I^a (Continued)

RPS1a	G A C U U A A U U C U A A G A A A A G U C A A G A U C U C G A G A C U A G C A A U A A C A A A A U G
RPS1b	G U A U A A C C U A G A G A A G A A U A A U A G A U A A A G A A A A A A G C A G A U A A A A A U G
RPL17a	G G G U C U U A C A A G C A A U A C A A A C C A A C A C A C C U A U A U A U A C U A A U A A U G
RPL25	U U C U U C U G C U G U U G A A A A G G C U A A A C A A A G A A G A U C A U C A U A A A G A U A A A U G
RPL29	----- A U U A A A U C C C A G A A C A A U A U A U C C A A C U A A U C A A G A A U G
RPL46	----- A A G C A A A U A A A C A C A G A U A G A U C A A A C A U A U G
RPS24	----- A A G A C C A A C A U A G C A U A U C C A A A G C A U A U C A A A G A U G
RPS33	U G U U C G U U U U C G A U U C U U C U C A A A A G U A G A A A C C A A G C U A G C A U A U C A U C A U G
SIR2g	G U A G A C A C A U U C A A A C C A U U U U U C C U A U C G G C A C A U U A A A G C U G G A U G
SIR3g	C A G A G G U U U A A G A A A G U U G U U U U G U U C A A C A A U U G G A U U A G C U A A A U G
SPT2	G A G A G G G A C A G G G A C U U G A G U C C U A U U C A A G U G A U A U U U U A G U U A U G
SUC2	----- C A A G C A A A C A A A A A G C U U U C U U U C A C U A A C G U A U A U G A U G
SUC7	A G U G A A C A A G C A A A A C A A A A A G C U U U C C U U U C A C U A A C G U A U A U G A U G
TOP1	A A A A A U C U A A A G G G A G G G C A G A G C U G A A C C U U G A A A C G C U G U A A A A U G
TRP1	U G A G C A C G U G A U A U C G U G A U A A U C C A C A C A A A G C A G C A G C U G G A A G U A U G
TRP2	A C A A U C G A U A A U A G C A C U G A U A U U C G A U U G G A A A A A G G C A A A A A U G
TRP3	C C A A A U C U G U U U G G U C U C A U A A G A A C G C C A U A A A A G A A A A A A U G
TRP5	----- A A A C U C G A U U G C U C C A A A A A G G G A C A U A G C A C A C G A C A A U G
TUBb	A A C G A G C U A C C A C C U A C A A A A G C A A A A U C C A C A A A A G C A U A U A U G
YP2onc	U A C U G U A A G G C C A C G U U G A A A A U A A A A A A C A A A A G C A U A A U A U G
Percent A:	32 25 26 34 28 32 30 30 31 41 35 38 33 45 40 31 36 35 33 48 39 50 46 42 43 46 33 50 49 48 45 50 39 54 48 54 45 40 55 45 50 36 49 50 81 51 63 1 0 0 0 0
Percent U:	25 27 20 23 19 21 32 26 26 21 19 22 16 24 27 29 26 22 19 20 21 23 16 21 22 27 18 16 20 26 21 23 20 24 21 21 28 22 22 21 38 22 03 24 09 00 1 0 0 0
Percent G:	11 11 13 09 34 11 06 08 09 09 19 14 20 09 14 14 05 15 16 13 09 08 11 17 08 15 10 13 13 14 09 14 15 08 07 09 11 15 06 10 15 14 10 07 09 08 00 0 1 0 0
Percent C:	06 13 18 11 20 16 13 19 20 10 13 19 16 21 15 20 20 16 22 14 25 15 14 20 22 16 27 18 21 18 19 15 23 18 21 16 23 18 17 23 15 13 19 21 06 17 19 00 0 0 0 0

Alignment of the nucleotides preceding the startcodon of 96 yeast genes from Genbank dated August 4, 1986.

The "RNA", strand is shown in this and all other tables. All 96 sequences were arranged alphabetically and aligned at the startcodon shown to the right. The frequency of occurrence of each base at each position is shown. The nucleotides that occur at a frequency of 51 percent or greater are boxed. Throughout the paper the first base of the startcodon is numbered 1. The full name of each gene shown here is given in Table Ic. Several genes have fewer than 48 bases of untranslated sequence. In these cases the most leftward base shown is the 5' end of the mRNA.

Table I^b Nucleotide sequences following the start codon

ACT	A U G G A U U C U G A G G U U G C U G C U U U G G U A U U G A U A A C G G U U C U G G U A U G U G
ADE4	A U G U G U G G U A U U U U A G G U A U U G A U U A G C A A A C C A A A C C A C U C C A G U A G C
ADH1	A U G U C U A U C C C A G A A C C U C A A A A A G G U G U U A U C U U C U C A G A A U C C C A C G G
ADH3	A U G U U G A G A C G U C A A C A U U G U U C C A A G C G U G U C C A C C A A G C C U A U U
ADR2	A U G U C U A U U C C A G A A C C U C A A A A A G C C A U U A U C U U C U C A G A A U C C A A C G G
ARG4	A U G U C A G A C G G C A C U C A A A A C U A G U G G G U G G G A G A U U C A C U G G U G A A A C
CCS	A U G U C A G C G A U A U A U C A A C A C A U A G C A A A A G U G A U U C U U A U C A A G G G G C U C
CDC28	A U G A G C G G U G A U U U A G C A A U U A C A A A A G A C A U G A G A A G C U G A A G C G G U A G G
CDC8	A U G A U B G G U C G U G C A A A U U A U A C U G A U A G A A G G A U U G G A U U G G A C U G G
CPA1	A U G U C U C C G C U G G C A C A A A A G C U A C U U U C U G U A U U C A A A A U G G C U U C
CPA2	A U G A C A U C G A U U U A U A C A U C A A C A G A G C C U A C G A A U U C U G C U U U A C U A C
CPAX	A U G U C U C C G C U G C A C A A A A G C U A C U U U C U G U A U C C A A A U U G G U C C U C
CPB1	A U G U U U U A C C U C G U C G U U C G G U A C A G G A C C G A G A G G U U U A A A A A U
CUP1	A U G U U C A G C G A U U A A U U A A C U U C C A A A U G A A G G C U A U G A G U G C C A A G
CYC	A U G U U U C A A A U C U A A C G U U A A C G U U G G C C A A A G G A C C C U C G A A A A G
CYC1	A U G A C U G A A U U C A A G G C G G U U C U G C U A G A A A A G G U G C U A C A C U U U U C A A
CYC17	A U G A C A U G U U G G A C U A G U U G G U G A U A C U G G G A C A C A C A A A A G A U A A C
CYC1x	A U G A C U G A A U U C A A G G C C G G U U C U G C U A G A A A A A G G U G C U A C A C U U U U C A A
CYC4	A U G C U U C A G U C A U C G U C A A U C U A U A G A U U U U C A A G C C A G C C A C A A G A A C
CYC7	A U G G C U A A A G A A A G U A C G G G A U U C A A C A G G C C U G C A A A A A G G G U G C
CYCr	A U G C C A C A G U G U U U A C G U C U A U U A C G A U U U G G A C U A U A U U U U G A A
EF1a	A U G G U U A A A G A G A A G U C U C A C A U U A C G U U G U C G U U A U C G G U C A U G U C G A
EF1.b	A U G G U U A A A G A G A A G U C U C A C A U U A C G U U G U C G U U A U C G G U C A U G U C G A
ENOa	A U G G C U G U C U C U A A A G U U U A C G C U A G A U C C G U C U A C G A C U C C G U G G U A A
ENOb	A U G G C U G U C U C U A A A G U U U A C G C U A G A U C C G U C U A C G A C U C C G U G G U A A
G3PDa	A U G G U U A G A G U U G C U A U U A C G U U U C G G U A G A A C G G U A G A U U G G U C A U
G3PDb	A U G G U U A G A G U U G C U A U U A C G G U U C G G U A G A N U C G G U A G A U U G G U U U
G3PDc	A U G A U C A G A A U U G C U A U U A C G G U U C G G U A G A A C G G U A G A U U U G G U U U
GAL	A U G A C U A A A U C U C A U U C A G A A G A G U A U U G A C C U G A G U U C A U U U C U A G
GAL10	A U G A C A G C U C A G U U A C A A A G U G A A G U A C U U C U A A A U U G U U U G G U U A C
GAL1p	A U G A C U G C U G A G A A U U G A U U U U C U A G C C A U U C C A U U G C A G U U A C A A
GAL4	A U G A A G C U A C U G U C U U C U A C G A A C A A G C A U G C G A U A U U U G C C G A C U U A A

Table I^b (Continued)

GAL80g	AUGGA	CUACA	CAAG	GAG	AUCU	UCGG	UCUCA	ACCG	UGCCU	AAUG	CAGC	UCC
CCN4	AUGU	CGAA	UAUC	AGCC	AAAG	UUUA	UUUG	CUUU	AAAU	CCAA	UGGG	UUUC
CDHm	AUGU	CGAG	GCAG	AAUU	UCAAC	AAAG	CUUA	CGAA	GAAG	UUUG	UCCU	UUU
H2A1	AUGU	CGG	UGGU	AAAG	UGGU	AAAG	CGUUC	UCCAG	UGCU	AAAG	CUUCA	
H2B1	AUGU	CGU	GCUA	AAAG	CCGAA	AAAG	CCAG	CCUUC	CAAG	CCCG	AGCU	GA
H2B2	AUGU	CGC	UCUG	CCG	CGAA	AAAG	AAAC	CAGC	UUC	CAAG	CCU	CAG
H3cI	AUGG	CGCA	GAACA	AAAG	CAAC	CAGC	AGAA	AAUCC	ACUG	GUU	GAAG	CC
H3cII	AUGG	CGCA	GAAC	UAAC	CAAC	CAGC	UJAG	AAAU	CCAC	UGGU	GUAA	AG
H4cI	AUGU	CGG	UAG	AGGU	AAAG	GGUU	AAAG	GGUUA	AGGU	AAAG	GGU	GC
H4cII	AUGU	CGG	UAG	AGGU	AAAG	GGUU	AAAG	GGUUA	AGGU	AAAG	GGU	GC
HIS1	AUGGA	UUG	CGU	AAAC	CAUC	UAAC	CGA	UAG	CAU	CGU	UU	GC
HIS3	AUGA	CGA	GAG	CAAA	AGCC	UAAG	CUAA	AGCU	AAU	CAAA	UGAA	CC
HIS4	AUGG	UUG	UUG	CCG	AAUUC	UAC	CGUU	AAU	UGA	UGA	UUC	GG
HMLa11	AUGU	UUU	ACU	UGAA	GC	UC	GUU	CA	AAU	UA	AG	AA
HMLa12	AUGA	AAU	AAAU	CA	CC	AAU	UA	AG	CA	UU	AA	AG
HSP90	AUGG	GUU	AGU	AAAC	UUU	UGA	UUUC	AGC	UGA	AAU	UAC	UG
HXK1	AUGG	UUU	CAU	UU	AG	GUCC	AA	AG	AA	CC	CA	AG
HXK2	AUGG	UUU	CAU	UU	AG	GUCC	AA	AG	AA	CC	CA	AG
LEU1	AUGG	UUU	CAU	UU	AG	GUCC	AA	AG	AA	CC	CA	AG
LEU2	AUGU	CGU	CCU	UAG	UC	UG	CCU	UA	AG	GA	UAG	GU
MATa11	AUGU	UUU	ACU	UG	AA	GC	UU	UA	AA	GA	CA	AA
MATa12	AUGA	AAU	AAU	UA	CC	AAU	UA	AG	CA	UU	AA	AG
MES1	AUGU	CUU	UUU	CCU	CAU	UU	CCU	UU	UA	AA	UAA	AG
MFA	AUGA	AAU	UU	CCU	U	CA	UUU	U	AC	UG	CA	UU
MFA1g	AUGA	GAU	UU	CCU	U	CA	UUU	U	AC	UG	CA	UU
MFA2g	AUGA	AAU	UU	CAU	U	CCU	UU	U	CA	UUU	U	AG
ODCd	AUGU	CGA	AAU	UC	UA	AA	GG	AA	CC	UG	CU	CA
ODCf	AUGU	CGA	AAU	UC	UA	AA	GG	AA	CC	UG	CU	CA
PGK	AUGU	CUU	AAU	UC	UA	AA	GG	AA	CC	UG	CU	CA
PHO3	AUGU	UUU	AAU	UC	UA	AA	GG	AA	CC	UG	CU	CA
PHO5	AUGU	UUU	AAU	UC	UA	AA	GG	AA	CC	UG	CU	CA
POR	AUGU	CUU	CCU	CA	UU	U	AC	AG	CG	UA	U	CC
PPR1	AUGAA	CGA	AAU	AAU	UA	AG	CA	UCC	AA	AA	AG	UA
PPR2	AUGAA	CGA	AAU	AAU	UA	AG	CA	UCC	AA	AA	AG	UA
PUT2	AUGC	UAU	CAG	CA	AG	GU	CCU	CA	AA	CU	UA	UA
PYK	AUGU	CUU	AG	AAU	U	AA	GA	U	AC	CU	AA	U
RAD1	AUGU	CUU	CAG	UAU	UU	U	AC	GG	GC	AG	CU	GA
RAD2g	AUGG	GUU	GGU	CAU	UUU	UG	GGU	AAU	UG	CG	UU	CA
RAD3	AUGA	AGU	UUU	U	AA	UG	AAU	U	CC	AG	UU	CA
RAD6	AUGU	CGA	CAU	CAU	AG	AA	GG	AAU	UG	AAU	U	CA
RASH1r	AUGCA	GGG	AAU	AAU	U	CA	AAU	AA	GG	AAU	U	CA
RASH2r	AUGC	GUU	UGA	AA	CA	AG	U	CA	AAU	AA	GG	AAU
RP13	AUGU	CUU	CA	CA	AAU	U	CA	AAU	AA	GG	AAU	U
RP29	AUGA	AGU	UGA	AAU	U	CA	AAU	AA	GG	AAU	U	CA
RP51a	AUGG	GUU	AGU	U	AA	CC	AA	AG	CC	GU	CA	AA
RP51b	AUGG	GUU	AGU	U	AA	CC	AA	AG	CC	GU	CA	AA
RPL17a	AUGU	CGU	GU	AA	CC	GU	CG	U	CA	AA	GU	AA
RPL25	AUGG	CUU	CCA	UC	GU	CA	AA	GG	CU	AA	GA	AA
RPL29	AUGCC	UUC	CA	AAU	U	CA	AAU	U	CA	AAU	U	CA
RPL46	AUGG	CUU	CCU	CA	AAU	U	CA	AAU	U	CA	AAU	U
RPS24	AUGA	CGA	CAU	U	CCU	U	U	AG	CGU	AAU	U	CA
RPS33	AUGGA	UAA	CA	AA	CC	CA	U	U	AG	CGU	AAU	U
SIR2g	AUGA	CGA	CAU	U	CCU	U	U	AG	CGU	AAU	U	CA
SIR3g	AUGG	CUU	AAA	CAU	U	U	U	GG	CA	AAU	U	CA
SPT2	AUGA	GUU	UUU	CUU	CC	AAU	U	U	CA	AAU	U	CA
SUC2	AUGC	UUU	U	GC	AA	GU	U	U	U	U	U	U
SUC7	AUGC	UUU	U	GC	AA	GU	U	U	U	U	U	U
TOP1	AUGA	CUU	AAU	UC	UG	U	CC	AA	GU	AAU	U	CA
TRP1	AUGU	CUU	GUU	AAU	U	U	U	CA	AG	GU	AAU	U
TRP2	AUGA	CGU	GUU	CC	AAU	U	U	CA	AG	GU	AAU	U
TRP3	AUGU	CUU	GUU	CC	AAU	U	U	CA	AG	GU	AAU	U
TRP5	AUGU	CGA	GA	CA	U	U	U	U	U	U	U	U
TubB	AUGA	GGA	AAU	CAU	U	U	U	U	U	U	U	U
YP2onc	AUGA	AAU	AG	CAU	U	U	U	U	U	U	U	U

Percent A:	10	00	28	16	14	35	30	36	27	29	35	39	34	33	35	27	33	41	45	29	32	36	33	34	36	35	27	27	32	33	25	39	36	34	28	33	24	21	25	36	28	21	18	29	32	24	40	27		
Percent U:	00	10	00	38	21	57	25	26	29	22	27	40	21	23	35	27	36	44	34	34	35	27	36	36	17	36	30	39	43	22	25	34	15	27	38	20	26	42	20	26	29	22	42	20	35	35	22	34		
Percent G:	00	00	10	26	11	09	30	22	17	26	07	14	30	13	16	21	07	06	21	13	16	32	15	08	27	17	11	36	19	18	28	21	13	34	17	15	33	10	14	36	16	11	34	25	19	36	08	15	33	20
Percent C:	00	00	00	08	52	20	09	22	18	25	36	11	10	30	16	17	40	17	15	19	30	08	31	22	21	31	22	08	25	13	18	21	38	11	20	24	19	30	21	13	33	23	16	32	22	15	26	36	05	29

Alignment of the 47 nucleotides following the startcodon. The same genes as shown in Table Ia are aligned in the same order. In this table nucleotides occurring at a frequency of 51 percent or greater are boxed.

Table 1^c Name and reference of genes listed in table 1^a and 1^b.

ACT	actin, Gallwitz (1980) <i>PNAS</i> 77, 2546; Ng (1980) <i>PNAS</i> 77, 3912 (1980); Domdey (1984) <i>Cell</i> 39, 611; Nellen (1981) <i>J Mol Appl Genet</i> 1, 239
ADE4	amidophosphoribosyltransferase, Maentsaelae (1984) <i>JBC</i> 259, 8478
ADH1	alcohol dehydrogenase, Bennetzen (1982) <i>JBC</i> 257, 3018
ADH3	alcohol dehydrogenase III, Pilgrim (1985) Unpublished, Biochem Dept, U Washington, Seattle WA 98195
ADR2	alcohol dehydrogenase II Russell (1983) <i>JBC</i> 258, 2674
ARG4	argininosuccinate lyase, Beacham (1984) <i>Gene</i> 29, 271
CCS	citrate synthase, Suissa (1984) <i>EMBO J</i> 3, 1773
CDC28	cell division control protein, Loerincz(1984) <i>Nature</i> 307, 183
CDC8	CDC8 gene, Birkenmeyer (1984), <i>Mol Cell Biol</i> 4, 583
CPA1	carbamyl phosphate synthetase small subunit, Nyunoya (1984) <i>JBC</i> 259, 9790
CPA2	arginine-specific carbamyl phosphate synthetase large subunit, Lusty (1983) <i>JBC</i> 258, 14466
CPAx	carbamoyl-phosphate synthetase small subunit, Werner (1985) <i>Eur J Biochem</i> 146, 371
CPB1	CBP1 gene, Dieckmann (1984) <i>JBC</i> 259, 4732
CUP1	copper chelatin, Karin (1984) <i>PNAS</i> 81, 337; Butt (1984) <i>PNAS</i> 81, 3332
CYC	cytochrome c1 precursor (nuclear), Sadler (1984) <i>EMBO J</i> 3, 2137
CYC1	CYC1 gene promoter region, McNeil (1985) <i>Mol Cell Biol</i> 5, 3545
CYC17	17-kd subunit of ubiquinol-cytochrome c reductase (nuclear), Van Loon (1984) <i>EMBO J</i> 3, 1039
CYC1x	iso-1-cytochrome c, Smith (1979) <i>Cell</i> 16, 753-761; Boss (1981) <i>JBC</i> 256, 12958
CYC4	cytochrome c oxidase subunit IV, Maarse (1984) <i>EMBO J</i> 3, 2831
CYC7	iso-2-cytochrome c, Montgomery (1980) <i>PNAS</i> 77, 541-545; Montgomery (1982) <i>JBC</i> 257, 7756
CYCr	ubiquinol-cytochrome c reductase 14 kd subunit, De Haan (1984) <i>Eur J Biochem</i> 138, 169
EF1a	elongation factor EF-1 alpha (TEF1), Schirmaier (1984) <i>EMBO J</i> 3, 3311
EF1ab	EF-1-alpha* (elongation factor 1-alpha), Cottrelle (1985) <i>JBC</i> 260, 3090
ENOa	enolase (clone peno46), Holland (1981) <i>JBC</i> 256, 1385; Holland (1983) <i>JBC</i> 258, 5291
ENOb	enolase (clone peno8), <i>ibid</i>
G3PDa	glyceraldehyde-3-phosphate dehydrogenase, Holland (1979) <i>JBC</i> 254, 9839
G3PDb	glyceraldehyde-3-phosphate dehydrogenase, Holland (1980) <i>JBC</i> 255, 2596
G3PDc	glyceraldehyde-3-phosphate dehydrogenase, Holland (1983) <i>JBC</i> 258, 5291
GAL1	GAL1 inducible promoter, Johnston (1984) <i>Mol Cell Biol</i> 4, 1440
GAL10	GAL10 inducible promoter, <i>ibid</i>
GAL1p	GAL7 gene, transcript initiation region, Nogi (1983) <i>NAR</i> 11, 8555
GAL4	positive regulator of galactose inducible genes, Laughon (1984) <i>Mol Cell Biol</i> 4, 260
GAL80	GAL80 regulatory gene, Nogi (1984) <i>NAR</i> 12, 9287
GCN4	GCN4 gene, Hinnebusch (1984) <i>PNAS</i> 81, 6442
GDHm	NADPH-dependent glutamate dehydrogenase, Moye (1985) <i>JBC</i> 260, 8502
H2A1	histone h2a-1, Choe (1982) <i>PNAS</i> 79, 1484
H2B1	histone H2B-1, Wallis (1983) <i>Cell</i> 22, 799
H2B2	histone H2B-2, <i>ibid</i> ; Wallis (1983) <i>Cell</i> 35, 711
H3cI	histone copy-I H3, Smith (1983) <i>JMB</i> 169, 663
H4cI	histone copy-I H3, <i>ibid</i>
H3cII	histone copy-II H3, <i>ibid</i>
H4cII	histone copy-II H3, <i>ibid</i>
HIS1	atp phosphoribosyltransferase, Hinnebusch (1983) <i>JBC</i> 258, 5238
HIS3	imidazoleglycerolphosphate dehydratase, Struhl (1981) <i>JMB</i> 152, 553
HIS4	HIS4 gene, Farabaugh (1980) <i>Nature</i> 286, 352-356; Donahue (1982) <i>Gene</i> 18, 47
HMLa1	mating-type locus HML-alpha-1, Nasmyth (1980) <i>Cold Spring Harb Symp Quant</i>

Table 1^c Continued

	<i>Biol</i> 45, 961; Astell (1981) <i>Cell</i> 27, 15
HMLa12	mating-type locus HML-alpha-2, <i>ibid</i>
HSP90	heat shock-inducible gene, Farrelly (1984) <i>JBC</i> 259, 5745
HXK1	hexokinase P-I, Kopetzki (1985) <i>Gene</i> 39, 95
HXK2	hexokinase P-II, Froehlich (1985) <i>Gene</i> 36, 105
LEU1	isopropylmalate-1 (IPM-1), Hsu (1984) <i>JBC</i> 259, 3714
LEU2	beta-isopropylmalate (beta-IPM) dehydrogenase, Andreadis (1982) <i>Cell</i> 31, 319; Andreadis (1984) <i>JBC</i> 259, 8059
MATa11	mating-type locus MAT-alpha-1, Nasmyth (1980) <i>Cold Spring Harb Symp Quant Biol</i> 45, 961; Tatchell (1981) <i>Cell</i> 27, 25
MATa12	mating-type locus MAT-alpha-2, <i>ibid</i>
MES1	methionyl-trna synthetase, Walter (1983) <i>PNAS</i> 80, 2437
MFA	pheromone MF-alpha, Kurjan (1982) <i>Cell</i> 30, 933
MFA1g	pheromone MF-alpha-1, Singh (1983) <i>NAR</i> 11, 4049
MFA2g	pheromone MF-alpha-2, <i>ibid</i>
ODCd	OMP decarboxylase, Rose (1984) <i>Gene</i> 29, 113
ODCf	OMP decarboxylase, <i>ibid</i>
PGK	3-phosphoglycerate kinase, Hitzeman (1982) <i>NAR</i> 10, 7791
PHO3	acid phosphatase, Bajwa (1984) <i>NAR</i> 12, 7721
PHO5	acid phosphatase, <i>ibid</i>
POR	porin, Mihara (1985) <i>EMBO J</i> 4, 769
PPR1	pyrimidine pathway regulatory 1 (PPR1) gene, Kammerer (1984) <i>JMB</i> 180, 239
PPR2	PPR2 gene, regulating dihydroorotase production, Hubert (1983) <i>EMBO J</i> 2, 2071
PUT2	P5C dehydrogenase, Krzywicki (1984) <i>Mol Cell Biol</i> 4, 2837
PYK	pyruvate kinase, Burke (1983) <i>JBC</i> 258, 2193
RAD1	RAD1 protein, Yang (1984) <i>Mol Cell Biol</i> 4, 2161
RAD2g	RAD2 protein, Nicolet (1985) <i>Gene</i> 36, 225
RAD3	RAD3 protein, Naumovski (1985) <i>Mol Cell Biol</i> 5, 17
RAD6	RAD6 protein, Reynolds (1985) <i>PNAS</i> 82, 168
RASH1r	ras-H related protein c-ras-sc-1, Dhar (1984) <i>NAR</i> 12, 3611
RASH2r	ras-H related protein c-ras-sc-2, <i>ibid</i>
RP13	ribosomal protein l3 (tcm1), Schultz (1983) <i>J Bacteriol</i> 155, 8
RP29	ribosomal protein 29, Mitra (1984) <i>JBC</i> 259, 9218
RP51a	ribosomal protein 51A, Teem (1983) <i>PNAS</i> 80, 4403
RP51b	ribosomal protein 51B, Abovich (1984) <i>Mol Cell Biol</i> 4, 1871
RPL17a	ribosomal protein L17a, Leer (1984) <i>NAR</i> 12, 6685
RPL25	ribosomal protein L25, <i>ibid</i>
RPL29	ribosomal protein L29, gene CYH2, Kaeufer (1983) <i>NAR</i> 11, 3123
RPL46	ribosomal protein L46, Leer (1985) <i>NAR</i> 13, 701
RPS24	ribosomal protein S24, <i>ibid</i>
RPS33	ribosomal protein S33, Leer (1983) <i>NAR</i> 11, 7759
SIR2g	silent information regulator protein, Shore (1984) <i>EMBO J</i> 3, 2817
SIR3g	silent information regulator protein, <i>ibid</i>
SPT2	SPT2 gene encoding regulatory protein, Roeder (1985) <i>Mol Cell Biol</i> 5, 1543
SUC2	invertase, Carlson (1983) <i>Mol Cell Biol</i> 3, 439
SUC7	invertase, Sarokin (1985) <i>NAR</i> 13, 6089
TOP1	topoisomerase I, Thrash (1985) <i>PNAS</i> 82, 4374
TRP1	trp1 (n-(5'-phosphoribosyl)-anthranilate, Tschumper (1980) <i>Gene</i> 10, 157
TRP2	anthranilate synthase, component I, Zalkin (1984) <i>JBC</i> 259, 3985
TRP3	anthranilate synthase, component II, <i>ibid</i>
TRP5	tryptophan synthase, Zalkin (1982) <i>JBC</i> 257, 1491
TUBb	beta-tubulin, Neff (1983) <i>Cell</i> 33, 211
YP2onc	YP2 protein proto-oncogene (human c-has/bas), Gallwitz (1983) <i>Nature</i> 306, 704

Abbreviation and name of the genes used in Table Ia and Ib and their references describing the DNA sequence.

Optimal AUG context in yeast and mammalian mRNAs

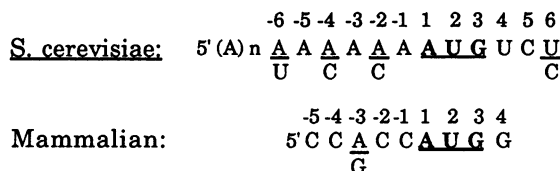


Figure 1. Comparison of the optional context around the start AUG derived for S. cerevisiae with that for mammalian mRNAs (5). With the exception of the A-residue at -3, the S. cerevisiae consensus is clearly different from that derived for mammalian mRNAs.

A strong bias exists for the first three bases downstream of the startcodon as well. Here, a U-residue prevails in the +4 and +6 positions (38 and 57 percent, respectively). G-residues are found at average frequency at +4 (G is the consensus nucleotide at position +4 in Kozak's rule; see figure 1). C-residues appear to be avoided (8 percent) at +4. Position +5 is occupied by a C-residue in 52 percent of all cases. Thus, the A/U/GCU type codons (Ser, Thr and Ala) are used most frequently with a preference for the UCU serine codon. The sequences further downstream have little or no biased nucleotide distribution.

The alphabetically listed sequences in Table I represent all translated mRNAs. Since the translation initiation efficiency is expected to be maximal in highly expressed genes, we analyzed a group of such genes separately. This group comprises the genes encoding glycolytic enzymes, ribosomal proteins, the elongation factor EF-1, and histones. Their gene products are abundantly produced in yeast cells and accordingly their coding sequence have a characteristic biased codon usage profile, i.e. they almost exclusively use 25 out of the 61 possible codons (14-16).

The sequences from the highly expressed genes are shown in Table IIa and IIb. Strikingly, the startcodon is preceded by an A-residue in almost all cases (with the exception of H4c1 and RP13). At position -2 an A-residue occurs in 50 percent of all cases. Position -3 is occupied by an A-residue in all cases. An A-residue occurs in 72 percent and 67 percent of all cases at positions -5 and -7, respectively. In all other positions of the untranslated region, A-residues prevail also strongly. The frequency of A-residues gradually decreases towards the 5' end. This biased nucleotide usage in the 5' untranslated region is more pronounced in the highly expressed genes shown in Table IIa than in all genes considered as a group

Table II^a Nucleotide sequences preceding the start codon of highly expressed genes

ACT	G C U U A C U G C U U U U U C U U C C C A G A G A U C G A A A A U U U U A C U G A U U A A C A A U G
ADR2	A A A A G C A U A C A U C A A C U A U C A A C U A U U A C U A U A U C G U A A U A C A A A U G
EF1ab	- - - - - - - - - - - - A A G C A U A G C A U C U A A U C U A G A G U U U U A U U A C A A A A U G
ENOa	- - - - - A A A C C C A A G A G C A A C U G C U U A U C A A C A C A A C A A C U A A U C A A A A U G
G3PDa	- - - A C A C C A A G A A C U A G U U U C G A A U A A A C A C A C A U A A A U A A U C A A A A U G
G3PDc	A A C A C A C A C A A A A A C A G A U A U C A C U A A A U U A C A C A C A A A U A A U C A A A A U G
H2A1	A A U A A C A A C U C A A A A C A A A C U A A U U C A A U A C A U A U A A A A A A A A A A U G
H2B1	A U C C U A U A U A G A C A G U C A A A C C A C A A U A A A C C U A U A C A C A C A C A A A U G
H2B2	U U C U G A U U G C U C U A U A C U C A A A C C A A C A C A C A C A C U A U A A A C A A A A U G
H3cI	U C U A G U A U A U A G A G A A A C A U C U A A C A U A A A U A U A A A A C G C A A A C A A U G
H4cI	A C U A A A G C A A C A A A C A A A A A C A A G C A A C A A A U A A U A A A A A A A A A U A U G
PGK	A U C A U C A A G G A A G U A U A U U A C U C U U U U A C A A C A A A A A U A U A A A C A A A U G
PYK	- - - - - - - - - - - G A C A C C A A U C A A A C A A A A A A A A A C A U C A U C A A A U G
RP13	U C G G U U U U G C A U C U C U A G A A C A A C A C A G U A U A C U A C A C A A U A U A A U G
RP29	- G C C A A G U U A G C G A A G A C A C C A A G A C A A U A C U U G A G A G U G A U A A A U G
RP51a	G A C U U A A U U C U A A G A A A A G U C A A G A U C U C G A A G A C U A G C A U A A C A A A U G
RPL17a	G G G C U U A C A A G G A U A C A A A C C A A C A C A C C A G U A U A U A U A A U A A A U G
RPL25	U U C U U C U G C U G U G A A A A G G C U A A C A C A C C A A G A A G A U C A A U A A A U G
Percent A:	33 22 06 39 28 50 22 39 28 33 39 56 28 44 61 61 50 39 39 50 33 44 61 39 56 50 44 56 56 61 56 44 33 50 50 50 44 50 61 67 33 72 39 10 50 89 10 00 00
Percent U:	22 22 22 28 28 17 39 28 17 22 22 17 28 11 22 11 22 33 11 22 17 22 17 11 33 22 28 17 11 22 22 22 39 44 28 22 22 28 22 17 33 22 22 00 17 06 00 10 00
Percent G:	17 11 11 06 17 00 11 11 17 06 17 11 17 17 11 06 06 11 22 11 00 00 06 17 06 00 00 06 06 11 00 06 06 06 00 06 06 11 06 06 11 00 00 00 00 00 00 00 10
Percent C:	00 22 39 11 17 22 17 11 28 28 11 06 22 22 06 22 22 17 28 17 50 33 17 33 28 17 33 11 22 22 17 17 28 22 06 17 22 22 17 11 06 22 06 39 00 33 06 00 00

Alignment of nucleotide sequences before the startcodon of highly expressed genes. Nucleotides occurring in more than 50 percent of all cases are boxed.

Table II^b Nucleotide sequences following the start codon of highly expressed genes

ACT	A U G G A U U C U G A G G U U G C U G C U U U G G U A U U G A U A A C G G U C U G G U A U G U
ADR2	A U G U C U A U U C C A G A A A C C A A A A A A G C C A U A U C U U A C G A A U C C A C G G
EF1ab	A U G G U A A A A G A G A A G A G U C U C A C A U A A C G U U G C G U U A C G G U C A A U C G A
ENOa	A U G G C U G U C U A A A G U U U A C G C U A G A U C C G U C A C G A C U C C G U G G U A A
G3PDa	A U G G U A G A G U U C A U A U A A C G G U U C G G U A G A U C G U A G A U U G U C U U
G3PDc	A U G A U C A G A A U U G C U A U U A A C G G U U U C G G U A G A A U C G U A G A U U G U C U U
H2A1	A U G U C C G G U G G U A A A G G U G U A A A G C U G G U C C A C C U A A A G C U C U C U
H2B1	A U G U C U G C U A A A G C C G A A A A G A A A C C A G C C U C C A A A G C C C A G C U G A A A A
H2B2	A U G U C C U C U G C G C C G G A A A A G A A A C C A G C U U C C A A A G C U C C A U G A A A A
H3cI	A U G G C C A G A A C A A A G C A A A C A G C A A A G A A A G U C C A U C G U G U A A G C C C C
H4cI	A U G U C C G G U A G A C U G U A A A G G U G U A A A G G U C U A G G U A A A G G U G U G C C A A
PGK	A U G U C U U U A U C U C A A A G U U G U C U G U C C A A G A U U U G A C U G A A A G A G A C A A
PYK	A U G U C U A G A U A A G A A A G A U U G A C C U C A U A A A C G U U G U U G C U G U G U U C U G A
RP13	A U G U C U C A C A G A A A G U A C G A A G C A C C A C G U C A C G G U C A U U A G G U U C U U
RP29	A U G A A G G U U G A A A U C G A U U C U U U U C A C G G U G C C A A A U C U A C C C A G C A G
RP51a	A U G G U A G A G U U A G A A C C A A G C C G U C A A G C G U C A A G C G U A U C U A A G C U U G A U
RPL17a	A U G U C C G G U A A C G G U G C U C A A G G U A C U A A G U U A G A A U C U C A U U A G G U C U
RPL25	A U G G C U C C A U C U G C U A A G G C U A C U G C C G C U A A G A A A C C U G U C G U A A G G G
Percent A:	10 00 00 11 11 00 39 11 44 33 28 39 39 33 44 44 28 33 56 22 44 22 33 28 11 44 28 22 11 28 28 22 50 33 28 22 28 06 22 22 44 11 17 11 17 28 11 44 44
Percent U:	00 10 00 50 11 61 17 22 44 22 22 39 06 11 33 11 17 50 22 11 28 17 17 50 22 28 17 11 22 61 28 28 22 17 33 39 11 22 56 33 17 28 22 22 61 22 33 22 17 28
Percent G:	00 00 10 39 11 06 33 44 00 39 17 11 56 17 17 39 11 11 28 11 28 39 22 06 33 11 00 50 33 17 28 17 06 33 17 06 61 22 00 33 28 11 50 28 22 61 17 17 22 22
Percent C:	00 00 00 00 67 33 11 22 11 06 33 11 00 33 17 06 28 11 17 22 22 00 39 11 17 50 39 11 22 11 17 28 50 00 17 28 06 28 39 11 33 17 17 33 06 00 22 50 17 06

Alignment of nucleotide sequences following the startcodon of the same highly expressed genes shown in Table IIa. Nucleotides occurring in more than 50 percent of all cases are boxed.

(Table Ia). G-residues are rare throughout the leader and are lacking in the five positions preceding the AUG. It is interesting to note that a string of ten G-residues just prior to the AUG startcodon has recently been shown to profoundly affect the expression of the Hepatitis virus coat protein gene in *S. cerevisiae*. When the G-residues preceding the AUG were

removed, the protein levels increased about 100-fold (17-18; Loren Schulz, personal communication).

After the startcodon of the highly expressed genes, the UCU and UCC serine codons also prevail as second codon at a frequency that is somewhat higher than that of all genes taken together. In Figure 1, the consensus sequence for mammalian mRNAs is compared with that from highly expressed yeast mRNAs.

DISCUSSION

The compilation presented here shows that AUG is used in all cases as protein initiation triplet. The question whether ribosomes of this yeast are absolutely restricted to the use of AUG for protein initiation is somewhat controversial. Sherman and Stewart (19) reported that no iso-1-cytochrome c is made when the starting AUG was mutated to GUG, AUA, CUG, AGG or AAG. However Zitomer et al. (20) studying fusions of the same CYC1 initiation region to the E. coli galactokinase gene, showed that the triplets AUA, UUG can be used at low efficiency provided that they are preceded by an A-residue at position -3 (20). No initiation occurred at AUA or UUG when position -3 was occupied by a U-residue.

Our compilation confirms the conclusion drawn earlier by Ammerer et al. (21) who compiled 20 initiation sequences. From the limited sequences available at that time, they also concluded that the untranslated region is rich in A-residues; that G-residues occur rarely in the 7 bases preceding the startcodon, and that position +6 is occupied frequently by a U-residue.

Although the nonrandom nature of the AUG context is clear, its purpose with respect to the mechanism of protein synthesis initiation is not. The extremely high frequency of A-residues in the 5' untranslated part suggests that absence of RNA structure is crucial for the scanning activity of the 40S subunit; the high A-content may prevent strong interaction between the leader and the rRNA within the subunit. It is possible that the distinct difference in nucleotide bias before and after the startcodon can be read by the subunit as a signal to terminate scanning and to initiate protein synthesis at the proper AUG.

The features of the translation initiation sites of yeast mRNAs are not only distinctly different from mammalian mRNAs but also from that of plant mRNAs (22). The untranslated region of plant mRNAs is A-U rich with a moderate preference for an A-residue at -3, but this A-residue is not flanked by C-residues as is the case in mammalian mRNAs. The preference of

Table III. Amino acid frequencies at the first 10 N-terminal positions

Amino Acid	Position									
	1	2	3	4	5	6	7	8	9	10
A	0	9	7	7	12	8	5	8	11	14
C	0	1	0	0	0	1	1	0	0	0
D	0	5	1	0	1	3	3	6	4	2
E	0	0	7	10	6	2	3	5	2	2
F	0	7	6	2	3	8	5	10	6	4
G	0	5	8	2	6	2	6	10	1	10
H	0	0	3	2	2	2	2	0	0	0
I	0	1	4	9	6	8	5	7	5	9
K	0	5	12	3	18	8	17	9	10	14
L	0	5	8	9	8	3	10	8	8	7
M	96	1	1	0	1	1	0	0	0	1
N	0	4	1	5	3	1	7	1	2	3
P	0	3	2	11	3	6	1	1	9	2
Q	0	1	3	6	1	5	6	2	4	3
R	0	3	10	5	5	4	1	4	10	7
S	0	29	12	12	10	12	9	9	7	8
T	0	11	3	5	5	12	6	12	6	1
V	0	6	6	6	4	5	3	2	8	7
W	0	0	0	0	0	0	1	0	2	0
Y	0	0	2	2	2	5	5	2	1	2

The frequency of occurrence of each of the first 10 amino acids of the mRNAs from Table I.

a G-residue at +4 coincides with Kozak's rule. This preference is accounted for by the extremely high frequency of occurrence of an Alanine codon as second triplet in the plant mRNA sequenced thus far (see below).

The prevalence of the UCU/C codons at the second position in yeast mRNAs may have a few interesting possible explanations. Bachmair *et al.* (23) showed that the N-terminal amino acid of β -galactosidase determines to a great extent its stability in yeast. They infer that this is likely to be true for all deblocked non-compartmentalized proteins. Thus, they divide the amino acids into a stabilizing and destabilizing group. The stabilizing amino acids are Met, Ser, Ala, Gly, Thr and Val. Table III shows that these amino acids occur in 63 percent of all cases at the second position. When the frequency of occurrence at the second position is compared with that at the following nine positions it appears that only serine (and possibly threonine) is used preferentially at the second position. This high frequency of serine usage may be related to its protein stabilizing role. It should be noted that the codons for the stabilizing amino acids Ser, Ala and Thr all have a C-residue in their central position and in most cases (i.e. the major codons of each family) have a C- or U-residue at the third position. This accounts for the occurrence of C- and U-residues at

positions +5 and +6 in the yeast consensus initiation site (Fig. 1) . The U-residue at the +4 position in the consensus sequence is accounted for by the high frequency of serine. It is interesting to note that plant mRNAs have in almost all cases, with three exceptions, at the second position also a codon for a stabilizing amino acid, namely alanine (see the compilation of 29 sequences by Heidecker and Messing, Ref. 23).

The high frequency of the serine codons UCU and UCC at the second position of yeast mRNAs may also play a role in mRNA translatability. This is the case for *E. coli* mRNAs. A mutational analysis of the second codon in the *lacZ* mRNA of *E. coli* showed that the nature of the second codon affects expression over a 20-fold range (24,25). In *E. coli*, serine is the second most frequently used amino acid for the second position (Ala is used most frequently, followed by serine, followed by lysine). Therefore, it is possible that the frequent use of the serine codon at the second position might likewise play a role in determining the translation initiation frequency in yeast mRNAs. Whether this is indeed the case remains to be proven experimentally. Other factors related to enzymatic requirements for the removal of the N-terminal methionine by methionine amino peptidase may also play a role in second amino acid selection (26).

ACKNOWLEDGEMENTS

We like to thank Dr. R. Hitzeman for helpful suggestions and comments. We thank Socorro Cuisia for typing the drafts and the final manuscript.

REFERENCES

1. Kozak, M. (1984) *Nucleic Acids Research* 12, 857-872.
2. Kozak, M. (1983) *Microbiol. Rev.* 47, 1-45.
3. Kozak, M. (1981) *Curr. Top. Microbiol. Immunol.* 93, 81-123.
4. Kozak, M. (1980) *Cell* 22, 7-8.
5. Kozak, M. (1984) *Nucleic Acids Research* 12, 3873-3893.
6. Kozak, M. (1986) *Cell* 44, 283-92.
7. Kozak, M. (1984) *Nature* 308, 241-246.
8. Kozak, M. (1981) *Nucleic Acids Research* 9, 5233-5262.
9. Liu, C.C., Simonsen, C.C. and Levinson, A.D. (1984) *Nature* 309, 82-85.
10. McNeil, J.B. and Smith, M. (1985) *Molecular and Cellular Biology* 5, 3545-3551.
11. Hsu, Y.P. and Schimmel, P. (1984) *J. Biol. Chem.* 259, 3714-3719.
12. Johnston, M. and Davis, R.W. (1984) *Molecular and Cellular Biology* 4, 1440-1448.
13. Zalkin, H., Paluh, J.L., van Cleemput, M., Moye, W.S. and Yanofsky, C. (1984) *J. Biol. Chem.* 259, 3985-3992.
14. de Boer, H.A. and Kastelein, R. (1986) in *Maximizing Gene Expression*, pp. 225-285, J. Davies, W. Reznikoff, and L. Gold, Eds., Butterworths Publishing Co., Boston.
15. Bennetzen and Hall (1982) *J. Mol. Biol.* 158, 573-597.

16. Ikemura (1982) *J. Mol. Biol.* 158, 573-597.
17. Kniskern, P.J., Hagopian, A., Montgomery, D.L., Burke, P., Dunn, N.R., Hoffman, K.J., Miller, W.J. and Ellis, R.W. (1986) *Gene* 46, 135-141.
18. Valenzuela, P., Gray, P., Quiroya, M., Zaldivar, J., Goodman, H.M. and Rutter, W.J. (1980) *Nature* 280, 815-819.
19. Sherman, F. and Stewart, J. W. (1982) In: *The molecular biology of the yeast Saccharomyces, Metabolism and Gene Expression*. Eds. J. N. Strathern, E. W. Jones and J. R. Broach, pp. 301-333.
20. Zitomer, R.S. et al. (1984) *Molecular and Cellular Biology* 4, 1191-1197.
21. Ammerer, G., Hitzeman, R., Hagie, F., Barta, A. and Hall, B.D. (1981) In: *Recombinant DNA. Proceeding of the third Cleveland Symposium on Macro molecules*, Cleveland, Ohio. Elsevier Scientific Publishing Co., Amsterdam, The Netherlands, pp. 185-197.
22. Heidecker, G. and Messing, J. (1986) *Ann. Rev. Plant Physiol.* 37, 439-466.
23. Bachmair, A. et al. (1986) *Science* 234, 179-186.
24. Looman, C., van Knippenberg, P.H. and de Boer, H.A. (1986) Ph.D. Thesis, pp. 70-80, University of Leiden, The Netherlands.
25. Looman, C., van Knippenberg, P.H., Eaton, D., Jhurani, P. and de Boer, H.A. *EMBO J.* (in press).
26. Stewart, J. W. et al. (1971) *J. Biol. Chem.* 246, 7429-7459.