The sequence flanking translational initiation site in protozoa

Kiyoshi Yamauchi
Department of Biology, Faculty of Science, Shizuoka University, 836 Oya, Shizuoka 422, Japan

Received January 24, 1991; Revised and Accepted April 12, 1991 GenBank accession no. M57542

ABSTRACT

Nucleotide sequences flanking the translational initiation site were compiled from protozoan nuclear genes and were compared in every protozoan group. The entire 5'-untranslated sequences were very rich in A- and T-residues, but poor in G- and C-residues in most protozoan genes except for the flagellated ones. The sequence $\overset{AAAAA}{TTTTT}A\overset{AAA}{TTT}AAAAATGAN\overset{A}{T}$ emerged as a consensus sequence flanking the initiation site in the major protozoan group, although the sequences upstream from −4 (four nucleotides upstream from the ATG codon) were divergent among ciliates, sarcodinians, and sporozoans. On the other hand, the consensus sequence for flagellates was revealed to be a simple feature. Only the nucleotide position −3 was occupied with a high frequency of A-residue, in other positions it appeared randomly. These facts suggest that the strong preference for A-residue at the position −3 is a universal feature in nuclear genes for all eukaryotes.

INTRODUCTION

A 5'-leader sequence of mRNA is well known to play an important role in protein biosynthesis. From the compilation and the analyses of the nucleotide sequences flanking the translational initiation site of nuclear genes, $CC\overset{A}{G}CCATG(G)$ emerged as the consensus sequence in vertebrates (1−4). Recently distinct consensus sequences were found in plants (5−6), *Drosophila* (7) and yeast (8). These were AACAATGGC, $\overset{C}{A}AA\overset{A}{C}ATG$ and $\overset{A}{T}A\overset{A}{C}A\overset{A}{C}AATGTC\overset{T}{C}$, respectively. A comparison of these consensus sequences suggested the possibility that such consensus sequences were divergent in every phyla, and that the A-residue at the position −3 was preferentially used in nuclear genes of all eukaryotes. To assess this possibility, I compiled and compared the sequences flanking the translational initiation site of protozoan nuclear genes. The comparison showed that the protozoan genes, except for flagellated ones, had a unique consensus sequence very rich in A- and T- residues. Flagellated protozoan genes have a simple consensus sequence with an A-residue occurring very frequently at position −3. It was a common feature not only in whole protozoans but also in all eukaryotes.

RESULTS AND DISCUSSION

Protozoan Consensus Sequence

Nucleotide sequences listed in the Appendix were collected from the data of complementary and/or genomic DNA sequences including the translational initiation site. A window at 16 bases around the initiation site was analyzed. Several sequences were excluded for which an ambiguity occured regarding the identification of the initiation site. The data sets for *Plasmodium*, *Tetrahymena* and *Triponosoma* contained a number of closely related genes. The sequences chosen for study were those in which the 5'-leaders showed the divergence in at least five positions. If a specific nucleotide was observed at a frequency greater than 50% at a specific position, it was defined as consensus nucleotide. If the sum of the frequencies of two nucleotides was greater than 75% and neither nucleotide met the criteria for a single consensus, they were assigned as co-consensus nucleotides.

The predicted initiation sites were not 5'-proximal ATG in 15 sequences out of 138 ones examined, although 7 sequences out of 15 were excluded because an ambiguity occured regarding the identification of the initiation site. From the frequency distribution of each nucleotide on the sequence surrounding the initiation codon for 131 protozoan nuclear genes, a consensus sequence was proposed according to the definition. However, when the frequency distributions were inspected in every taxonomic group, it was found that the consensus sequence at the positions −4 to −1 was almost the same for sarcodinians, ciliates, and sporozoans (Table 1), but was distinct from that of flagellates (Table 3). The mean values of the A- and T-residue content in the 5'-untranslated leaders were also estimated to be 84.9%, 84.8%, 80.3% and 63.5% for sarcodinians, ciliates, sporozoans and flagellates, respectively, as calculated from the sequence data in the Appendix. Judging from these two results, the whole population was subdivided into two groups: major protozoan group and flagellates. The 5'-leaders for each group were then analyzed.

For the major protozoan group as shown in Table 1, consensus sequences were obtained in three subgroups: AAAAAAAA-AAAAAATG$\overset{A}{G}$NA, $\overset{T}{A}A\overset{A}{T}ANAANT\overset{A}{T}AAAATG\overset{ACT}{GGA}$ and $\overset{AT}{TA}NTTNT\overset{T}{A}NAAAAATG\overset{AA}{GT}$, in sarcodinians, ciliates and sporozoans, respectively. However, the consensus sequences have some unreliability because they were derived from calculations using small subpopulations. The most conspicuous conserved feature among the three protozoan subgroups was the presence of A-residues with extremely high frequencies of 70−90% at

Table 1. Frequency of each nucleotide around the translational starting site in nuclear genes from the major protozoan group[+].

[++]	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	ATG	+4	+5	+6
Sarcodina																	
A:	25	22	30	24	24	24	24	25	27	26	40	36	35		17	16	24
G:	0	1	1	2	0	2	2	2	0	1	0	0	2		17	11	1
C:	2	5	3	4	3	6	3	0	5	4	1	2	3		3	10	6
T:	14	13	8	12	16	11	14	16	11	12	3	5	3		3	3	9
[+++]	A	A	A	A	A	A	A	A	A	A	A	A	A		A/G	N	A
Ciliata																	
A:	9	13	11	14	9	14	15	10	4	13	18	20	20		12	5	9
G:	0	1	1	0	1	0	1	4	2	0	2	1	2		10	9	1
C:	2	3	2	3	7	3	3	6	5	4	2	2	3		1	12	4
T:	11	5	8	5	5	7	6	5	14	8	4	3	1		3	0	12
	T/A	A	A/T	A	N	A	A	N	T	A/T	A	A	A		A/G	C/G	T/A
Sporozoa																	
A:	13	12	9	9	9	13	10	11	12	18	30	24	25		23	17	15
G:	3	2	3	4	2	7	1	5	4	3	0	6	5		5	8	9
C:	4	3	8	1	0	2	3	1	5	8	0	1	1		2	5	0
T:	12	15	12	18	21	10	18	16	12	4	4	3	3		4	4	10
	A/T	T/A	N	T	T	N	T	T/A	N	A	A	A	A		A	A/G	A/T
The sum of the three groups																	
A:	47	47	50	47	42	51	49	46	43	57	88	80	80		52	38	48
G:	3	4	5	6	3	9	4	11	6	4	2	7	9		32	28	11
C:	8	11	13	8	10	11	9	7	15	16	3	5	7		6	27	10
T:	37	33	28	35	42	28	38	37	37	24	11	11	7		10	7	31
	A/T	A/T	A/T	A/T	A/T	A	A/T	A/T	A/T	A	A	A	A		A	N	A/T

[+]: Data were calculated from the published papers in the Appendix.
[++]: Numbering begins with the A of the ATG codon as position +1; nucleotides 5′ to the site are assigned negative numbers.
[+++]: The nucleotides showing a frequency greater than 50% or the sum of the frequency greater than 75% of two nucleotide at a specific position are shown at the bottom of each table as consensus or co-consensus, respectively.

Table 2. Frequency of each nucleotide around non-functional ATG in 5′ leader sequences of nuclear genes from the major protozoan group.

	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	ATG	+4	+5	+6
A:	14	10	7	10	8	13	4	4	10	6	12	6	7		10	5	11
G:	6	4	6	7	6	6	4	6	4	4	5	3	6		4	8	7
C:	4	4	3	2	3	1	5	5	6	2	2	7	8		4	6	4
T:	1	7	9	6	9	6	13	11	6	14	7	10	5		8	7	4
	A	N	N	N	N	N	N	N	N	T	N	N	N		N	N	N

Data were calculated from the published papers marked by * in the Appendix. These sequences flanking the non-functional ATG were omitted. Numbering of the nucleotide positions and the criteria for consensus were carried out according to Table 1.

Table 3. Frequency of each nucleotide around the translational starting site in flagellated genes.

	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	ATG	+4	+5	+6
A:	10	12	6	4	13	6	10	9	6	12	23	9	7	5	2	7	
G:	5	4	8	4	1	7	2	4	3	4	1	1	11	4	7	3	
C:	9	4	8	9	5	7	6	3	9	7	3	12	5	8	16	12	
T:	3	7	5	10	8	7	9	12	10	5	1	6	5	11	3	6	
	N	N	N	N	A/T	N	N	N	N	N	A	N	N	N	C	N	

Data were calculated from the published papers in the Appendix.
Numbering of the nucleotide positions and the criteria for consensus were carried out according to Table 1.

the positions −3 to −1, although the compositions of A-residue were 49−65% in the window of the 5′-leader. Both G- and C-residues were found with low frequencies in this window. The compositions of G-residue were smallest values (2−11%) for all three subgroups. The sequences preceding the short stretch of A-residue diverged among the three subgroups. For sarcodinian, the consensus sequence showed a long homopolymer of A-residues at the positions −13 to −1. The ciliated consensus sequence possessed another short stretch of A-residue at the positions −12 to −7 and two T-residues on each side of the second A-stretch. Sporozoan exhibited a distinct bias in T-residue at the positions −13 to −5, suggesting that the sequence flanking

the initiation site could have a potential to form a stem structure by base pairing between the T-residue stretch and the A-residue stretch.

To clarify the specificity of the proposed consensus sequences, the frequency distribution of each nucleotide surrounding the 26 non-functional ATG triplets in the 5'-leader of the genes from the major protozoan group was summarized according to the same criteria described above and compared with those flanking the initiation sites. The composition of A-residue in the 13 positions preceding the non-functional ATG was 33% which was a lower value than those preceding the initiation site. As shown in Table 2, each nucleotide appeared to be randomly distributed at most positions of the sequences flanking the non-functional ATG in contrast to those flanking the initiation site. The preference for A-residues was not found at the position −3 to -1. These comparisons revealed that the consensus sequences proposed from the data in Table 1 were significantly specific to the positions surrounding the initiation site.

On the other hand, flagellated genes have a simple consensus sequence, which is $^{A}_{T}$NNNNNANNATGNC (Table 3). The highest occurrence of A-residue (82%) was found at the position −3 in this group. However, the frequency distribution of each nucleotide was approximately random except for the three positions: −9, −3 and +5.

This study could not confirm what feature the consensus sequences have in the highly expressed genes because of the small amount of sequence data in each subgroup.

Comparison of Consensus Sequences within Protozoan Groups

General views in consensus sequence between sarcodinians and ciliates were very similar, although the A-residue bias was stronger at the positions −13 to −4 in sarcodinians than in ciliates. The consensus sequence in sporozoans was similar to that in ciliates, but had a stronger T-residue bias at positions −12 to −6. Strong biases also existed for the first three bases (+4 to +6) downstream of the initiation site. Here, in addition to the A-residue biases for all of the three subgroups, the position +4 showed a high frequency of G-residue for both sarcodinians and ciliates and the position +5 and +6 frequently showed a G-residue and a T-residue, respectively, for both ciliates and sporozoans. These comparisons indicated that ciliates possessed a consensus sequence showing a middle feature between sarcodinians and sporozoans in the downstream as well as the upstream sequence. These comparisons indicated that the three subgroups might diverge from the same origin. Thus, the common feature of the consensus sequence could be proposed to be $^{AAAAA}_{TTTTT}A^{AAA}_{TTT}AAAAATGAN^{A}_{T}$ as shown from the sum of the requency distribution of the three subgroups (Table 1). The summary of the consensus sequence may prove useful for the identification of a putative initiation site for the major protozoan group.

Recently, the phylogenies of protozoans were examined using small-subunit rRNA genes (9) and large-subunit rRNA genes (10). This demonstrated that flagellated protozoans diverged from the mainstream of eukaryotic descent at a very early period and the major protozoan groups, ciliates, sarcodinians and sporozoans, emerged later, but before the metazoa-metaphyte-fungi radiation. The diversity of consensus sequences within protozoans examined in this study might well reflect an evolutionary distance among the taxonomic groups.

Comparison of Consensus Sequences among Protozoan and Other Eukaryotic Groups

Comparison of the consensus sequences for the major protozoan group with those for other eukaryotic groups revealed that A-residue bias in the sequence immediately preceding ATG was common to the sequences for yeast: $^{A}_{T}A^{A}_{C}A^{A}_{C}AATGTC^{T}_{C}$ (8) and for *Drosophila*: $^{C}_{A}AA^{A}_{C}ATG$ (7). The major difference in consensus between the major protozoan group and the above two groups was that the sequences for yeast and *Drosophila* have significant C-residue biases at the position −4, and −2 or −1. It was also notable that the C-residue biases were more remarkable in vertebrates: $CC^{A}_{G}CCATG$(G) (1−4) and plants: AACAATGGC (5,6). The repetition of G-residue for protozoan genes, which was noticed for vertebrate genes by Kozak (4), could not be detected. It is possible that the 5'-leader with high content of A- and T-residues plays a role specific for lower eukaryotes in translational efficiency. Mutzel *et al.* (11) noted that such a 5'-leader contained a signal that served for transcription or reinitiation of both transcription and translation. Also it has been suggested that non-histon proteins specific for the A- and the T-polymers play a role for gene packaging and/or gene activation in *Dictyostelium* (12). The effect of context on the translational and transcriptional initiation for protozoan genes remains to be clarified experimentally.

Only the A-residue at the position −3 was a common assignment in consensus sequences within protozoans examined in this study. The same position was also highly conserved in the consensus sequences for all investigated eukaryotic groups including vertebrates (1−4), plants (5,6), *Drosophila* (7) and yeast (8). This fact suggests that the A-residue at the position −3 has a most important role in recognition of AUG by translational machineries in all eukaryotic mRNAs. This conception also agreed with the experimental results, reported by Kozak (2), that purine residues have a potent effect upon translation rates at the position −3 where translation initiation is negatively affected by substitutions of non-consensus nucleotides. It is possible that the assignment of A-residue at position −3 must be read by the 40S ribosomal subunit as a signal to terminate scanning and to initiate protein synthesis at the proper ATG codon (13).

REFERENCES

1. Kozak, M. (1981) *Nucleic Acids Res.*, **9**, 5233−5252.
2. Kozak, M. (1986) *Cell*, **44**, 283−292.
3. Kozak, M. (1984) *Nucleic Acids Res.*, **12**, 857−872.
4. Kozak, M. (1987) *Nucleic Acids Res.*, **15**, 8125−8148.
5. Heidecker, G. (1986) *Ann. Rev. Plant Physiol.*, **37**, 439−466.
6. Lütcke, H. A., Chow, K. C., Mickel, F. S., Moss, K.A., Kern, H. F. and Scheele, G. A. (1987) *EMBO J.*, **6**, 43−48.
7. Cavener, D. R. (1987) *Nucleic Acids Res.*, **15**, 1353−1361.
8. Hamilton, R., Watanabe, C. K. and Herman, A. B. (1987) *Nucleic Acids Res.*, **15**, 3581−3593.
9. Sogin, M. L., Elwood, H. J. and Gunderson, J. H. (1986) *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 1383−1387.
10. Baroin, A., Perasso, R., Qu, L-H., Brugerolle, G., Bachellerie, J-P. and Adoutte,A. (1988) *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 3474−3478.
11. Mutzel, R., Lacombe, M.-L., Simon, M.-N., Gunzburg, J. and Veron, M. (1987) *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 6−10.
12. Garreau, H. and Williams, J.G. (1983) *Nucleic Acids Res.*, **11**, 8473−8484.
13. Kozak,M. (1989) *J. Cell. Biol.*, **108**, 229−241.

APPENDIX

SOURCE/GENE NAME	SEQUENCE[a]	START[b] DATA	A-T% OF[c] LEADER	REFERENCES[d]
SARCODINA				
Dictyostelium				
D. discoideum Actin B1:	-ATACAACAATAAAATGGAC	3,6	87	Firtel '79 PNAS 76, 6206.
2-SUB 1:	-ATAATATAATAAAATGGAT	2,6	–	"
2-SUB 2:	-TTTGAAATAATAAATGGAA	2,6	–	"
3:	-AAAAAAAAAAAAAATGGAA	2,6	–	"
5:	-TATATATAAAAAAATGGAC	2,6	–	"
CAP34 protein:	-ATAATAATATAAAATGGCC	2	–	Hartmann '89 JBC 264, 12639.
CAP32 protein:	-TATTTATTAAAAGATGACA	3,6	93	"
G protein 1:	-TATAAAATAAATAATGGGT	3,6	83	Pupillo '89 PNAS 86, 4892.
2:	-AAAAACTTAAAAAATGGGT	3,6	90	"
Myosin I heavy chain:	-AAAACAATTAAAAATGTCA	2,6 e	–	Jung '89 PNAS 86, 6186.
Myosin light chain:	-ACATAAAATAAAAATGGCC	3,5,6	78	Tafuri '89 MCB 9, 3073.
Ubiquitin (PCUB14):	-ATATTAATTAATTATGCAA	3,6	92	Ohmachi '89 Biochem 28, 5226
(PUCB17):	-ATAATTTTTTAATATGCAA	3,6	87	"
(PCUB19):	-ATAATTAATTATAATGCAA	3,6	95	"
Non-muscular alpha-actinin 3:	-ACATAAAAACAAAATGTCA	3,6	90	Noegel '87 FEBS 221, 391.
Dihydroorotate dehydrogenase:	-AAAAAAAAAAAAAATGGAA	2	–	Jacquet '85 Biochemie 67, 583.
Beta-N-acetylhexosaminidase A:	-AAAAAAAAAAAAAATGATC	3,4,6	82	Graham '88 JBC 263, 16823.
Spore coat protein SP60:	-AAAAAAAAAAAAAATGAAG	2,5	–	Fosnaugh '89 MCB, 9, 5215.
SP70:	-TACCAGTAATAAAATGAGA	2,5	–	"
Cell surface antigen PsA:	-ATAAATTAAAAAAATGAAA	3,	95	Early '88 MCB 8, 3458.
UDP glucose pyrophosphorylase:	-AAAAAAAAAAAAAATGACA	2,3,4	96	Ragheb '87 NAR 15, 3891.
109 gene protein 1:	-TCAAAATTCAAAAATGAAT	2,3,5	90	Giorda '89 JMB 205, 63.
2:	-AAAAAAATTAAAAATGAAT	2,5	–	"
cAMP inducible protein:	-AATAATATAAAAAATGAAT	2,4,5	90	Hopkinson '89 MCB 9, 4170
Cyclic nucleotide Phosphodiesterase:	-ATATACAAAAAAAATGGCA	1,3	83	Lacombe '86 JBC 261, 16811
Cysteine proteinase 2:	-TAATTATTTTAAAATGAGA	2,3,6	97	Pears '85 NAR 13, 8853.
DG17:	-TTTTTGTAAAAAAATGTCA	2,4	97	Driscoll '87 MCB 7, 4482.
Cell adhesion protein:	-AAATAAAATAAAAATGAAA	1,3	72	*Noegel '86 EMBO 5, 1473.
Discoidin-1 alpha:	-TCAACACAAATAAATG	4	87	Poole '84 JMB 172, 203.
beta:	-ACACACAATTAAAATG	4	86	"
gamma:	-TAAAATTTATAAAATG	4	92	"
Protein 253:	-TAAAATTGCAACCATGAAT	2,4	88	Ayres '87 MCB 7, 1823.
Protein 29C:	-TTTGTATTCCAAAATGAGA	2,4	85	"
Hisactophilin:	-TAATATAAATACAATGGGT	3,5	76	*Scheel '89 JBC 264, 2832.
M4 protein:	-TGGTTTGAATTTAATGAGA	2,4	81	*Kimmel '85 MCB 5, 2123.
Severin:	-AACTTTGACAAAAATGATT	1,3	88	Andre '88 JBC 263, 722.
cAMP dependent protein kinase:	-AAAAAAAAAAAAAATGACA	3,6	91	Mutzel '87 PNAS 84, 6.
Entamoeba				
E.histolytica Actin:	ATTCAATAAATATGGGA	4,6	91	Edman '87 PNAS 84, 3024.
Ferredoxin:	TCAAATCTAATGGGA	2,3,4,6	78	Huber '88 MBP 31, 27.
Naegleria				
N.gruberi Alpha-tubulin (1):	-ATAATAATACAAAATGAGA	3,6	68	Lai '88 JCB 106, 2035.
Pneumocystis				
P.carinii Thymidylate synthase:	-AATACTTTAAAACATGGTA	2,3,5	84	Edman '89 PNAS 86, 6503.
Acanthamoeba				
A.castellanii Actin:	-CTACATCATCAACATGGGA	2,4	63	Nellen '82 JMB 159, 1.
Non-muscle myosin heavy chain:	-CACATCAGCGAAGATGGCC	4,6	48	Hammer '87 JCB 105, 913.
CILIATA				
Paramecium				
P. caudatum Hemoglobin:	-AAAATGTCT	1,3 f	100	Yamauchi '91 unpublished
P. aurelia complex				
Surface antigen 156G:	TACTTTTAATGAAT	2,4	88	Prat '86 JMB 189, 47.
51C:	-TTTAATACTTAAAATGAAG	6	-	Preer 87' J. Protozool. 34, 418.
51H:	AAATAAAAATGCAA	2,4	100	"
51A:	-TTTTAATACTTTAATGAAT	2,4	93	Preer '87 Nature 314, 188.
Tetrahymena				
T. pyriformis Alpha-Tubulin:	-AGAAAAGTTAGAAATGAGA	2,4,6	75	Barahona '88 JMB 202, 365.
Beta-Tubulin I:	-AAAAGCAAATAAAATGAGA	2,4,6	85	"
II:	-TATCAAAATCAAGATGAGA	2,4,6	79	"
T. thermophila Actin:	-TAGATAAAGTAAAATGGCT	2,4,6	79	Cupples '86 PNAS 83, 5160.
Calcium-binding protein:	-AATAAATAAATAAATGGCT	3,	87	Takemasa '89 JBC 19293.
Ribosomal protein:	-AATAAAAATCAAAATGGGT	2,4,5	89	Nielsen '86 EMBO 5, 2711.
Histon H1:	-CAAATAATATAAAATGGCT	1	–	Wu '86 PNAS 83, 8674.
Histon H2B-1:	-AAATTCCATTTAAATGGCT	2,4	91	Nomoto '87 NAR 15, 5681.
Histon H3-I:	-CAAACTAATAAAAATGGCT	1	–	Horowitz '85 PNAS 82, 2452.
Histon H3-II:	-TACATAAGCAAAAATGGCT	1	–	"
Histon H4-I:	-AAAACTTACAAAAATGGCC	2,6	–	Bannon '84 NAR 12, 1961.
Histon H4-II:	-TAATCCAGCAAAAATGGCC	2,4	91	Horowitz '87 NAR 15, 141.

Stylonychia S. lemnae Alpha-1-tubulin:	-TCAACTCTTCATCATGAGA	2,4	80	Helftenbein '88 Curr Genet 13, 425.
Beta-1-tubulin:	-TCAAAAACAAACCATGAGA	2,4,6	80	Conzelmann '87 JMB 198, 643.
Beta-2-tubulin:	-AACTCAAGTCACAATGAGA	2,4,6	79	"
Euplotes				
E. raikovi Mating pheromone,Er-1:	-TATCAATTTTAGAATGAAC	1,3	67	*Micell '89 PNAS 86, 3016.
E. crassus Actin:	-ATATCTATCAAAAATGAGC	2,6	–	Harper '89 PNAS 86, 3252.
Beta-tubulin:	-TTTAAATCTAAAGATGAGA	2,6	–	"
Oxytricha				
O. nova C2 protein:	-TTACTATCGAGAAATGTCT	2,5	–	Klobutcher '84 Cell 36, 1045.
Actin:	ACTACACATGGCA	2,4,6	79	Greslin '88 DNA 7, 529.
O. fallax Actin:	-ACTACTCGTACATATGTCA	2,6	–	Kaine '82 Nature 295, 430.
SPOROZOA				
Plasmodium				
P.falciparum pf-Actin I:	-TATATCTGTAAAAATGGGA	3,6	87	Wesseling '88 MBP 27, 313.
Actin II:	-TCCTTTTCTTTAGATGTCT	3,6	72	*Wesseling '88 MBP 30, 143.
Hypoxanthin-guanine phosphoribosyl transferase:	-ATAATATTAGAAAATGCCA	3,6	92	King '87 NAR 15, 10469.
Major merozoite surface antigen:	-CTTTAATTCAATAATGAAG	3,6	94	Peterson '88 MBP 27, 291.
S antigen (FC27):	-AATATTATATACAATGAAT	2,6	–	Cowman '85 Cell 40, 775.
Blood stage antigen:	-TTCATATATCAAAATGAAG	2,5	–	Knapp '89 MBP 32, 73.
Surface antigen (pf12):	-AAAAGTATGATA	2,5	–	Elliott '90 PNAS 87, 6363.
Alpha-tubulinI:	-ATTTAAATAAAAAATGAGA	2,6	–	Holloway '89 M.Micro. 3, 1501.
Beta-tubulin:	-ATTTTATTAAAGAATGAGA	2,6	–	Delves '89 M.Micro. 3, 1511.
Multidrug resistance (MDR) protein:	-TTTGTGTTGAAAGATGGGT	2,6	–	Foote '89 Cell 57, 921.
Glutamic acid-rich protein:	-AACTTTTTAAAAAATGAAT	2	–	Triglia '88 MBP 31, 199.
Histidine-rich protein II:	-TTATTTAATAAAAATGGTT	2,3,4	100	Thomas '86 PNAS 83, 6065.
Glycophorin binding protein:	-ATTTTGTGTAAATATGCGA	2,5	–	Kochan '86 Cell 44, 689.
Thrombospondin related anonymous protein:	-AATTGTAAAAATAATGAAT	2,6	–	Kathryn '88 Nature 335, 79.
Glycoprotein 185:	-CTTTAATTCAATAATGAAG	3,6	87	Howard '86 Gene 46, 197.
Integral membrane protein (pf7):	-TATATTAGTCAAAATGAAG	2	–	Smythe '88 PNAS 85, 5195.
(Ag352):	-AAATTGTACAAAAATGAGA	2	–	Peterson '89 MCB 9, 3151.
Knop protein:	-TAATTATTAGAGAATGAAA	3	88	Kilejian '86 PNAS 83, 7938.
Serine-repeat antigen:	-CTCATATATCAAAATGAAG	3	88	Bzik '88 MBP 30, 279.
Distinct fast evolving repeats:	-AAAGAAAAGAAAAATGAAA	2	–	Scherf '88 EMBO 7, 1129.
Blood stage antigen:	-TAATATATTCAAAATGAAA	3,6	92	Coppel '85 PNAS 82, 5121.
P. knowlesi Circumsporozoite antigen:	-ATACAAGAACAAGATGAAG	2,5	72	Ozaki '83 Cell 34, 815.
P.malariae :	-GACTTGCTCCAACATGAAG	2,6	–	Lal '87 MBP 30, 291.
P.yoelii :	-AAAATGAAG	2,6	–	Lal '87 JBC 262, 2937.
P.berghei :	-GCATATATTTAAAATGAAG	2,6	–	Eichinger '86 MCB 6, 3965.
P.gallinaceum 25 Kd ookinete surface antigen:	-TCCTTTTTAAAAAATGAAT	3,6	94	Kaslow '89 MBP 33, 283.
P. lophurae Histidine-rich protein:	-AGGGAAAGGAAGTATGTTT	1,2,3,4	78	*Ravetch '84 Nature 312, 616.
Toxoplasma				
T.gondii Alpha-tubulin:	-CTTTTTCGACAAAATGAGA	2,4,6	55	Nagel '88 MBP 29, 261.
Beta-tubulin:	-TGCATCTTCCAAAATGAGA	2,6	–	"
28 Kd antigen:	-GTCTTGAAAGAGAATGTTT	3	52	*Prince '89 MBP 34, 3.
Theileria				
T.annulata 70kd heatshock protein:	-AATAGATTTAAAGATGACA	2,4,6	73	Mason '89 MBP 37, 27.
Babesia				
B. rodhaini surface antigen 26:	-TAGGTGTATATAGATGGCT	2,5	–	Snary '88 MBP 27, 303.
17:	-ATATTATATATAAATGTCA	2,5	–	"
B. bovis BabR:	-TTGAAGCTGTTGAATGGAA	3	61	*Cowman '84 Cell 37, 653.
FLAGELLATA				
Tripanosoma				
T. burucei				
Variant surface glycoprotein:	-GGAGCGACTCACAATGGAC	1,3	58	Boothroyd '82 JMB 157, 547.
Calmodulin:	-CACTTGATTTACGATGGCC	2,4	70	Tschudi '85 PNAS 82, 3998.
Disulphide isomerase like protein:	-AAAAAGCGTCACGATGCGC	2,4,5,6	73	Hsu '89 Biochem. 28, 6440.
Variant surface glycoprotein:	-AAAAATCGACCCGATGCTC	4	67	Scholler '88 MBP 29, 89.
Actin copy A:	-CTGCCATAAAATAATGTCG	2,4,6	67	Amer '88 MCB 8, 2166.
Fructose biphosphate aldolase:	-AACTGCAACGAAGATGTCC	3,6	62	Clayton '85 EMBO 4, 2997.
Expression site associated protein (ESAG 117a):	-TACTATATTGACAATGAAA	3,5	68	Cully '85 Cell 42, 173.
Glycosomal protein p60:	-CAATCATAACACAATGGCT	2	–	Kueng '89 JBC 264, 5203.
Phospholipase C:	-AAGAATCATTGTAATGTTT	3	61	Hereld '88 PNAS 85, 8914.
Heat shock protein 70:	-GCCTCTTTGAAGGATGACA	2,4	65	Glass '86 MCB 6, 4657.
Ornithine decarboxylase:	-CAGTAAGGCAAATATGACC	2,6	–	Phillips '87 JBC. 262, 8721.
Procyclic acidic repetitive protein:	-AGTAAAATTCACAATGGCA	3,4	73	Mowatt '87 MCB 7, 2837.
Phosphoglycerate kinase B:	-TCACAAATCAAATATGTCA	2,6	–	Osinga '85 EMBO 4, 3811.
C:	-ATTGTGATACAAGATGACC	2,6	–	"
RNA polymerase I:	-ACGCACAACTCCCATGTCG	2,4,6	50	Smith '89 JBC 264, 18091.
RNA polymerase II:	-AAGTAGTGCAAACATGTCA	2,4	57	Smith '89 Cell 56, 815.
RNA polymerase III:	-CGTTTCTTAACCGATGCTA	2,6	–	Köck '88 NAR 16, 8753.

Triosephosphate isomerase:	-ACCGTGCCAAATTATGTCC	2,4,6	67	Swinkels '86 EMBO 5, 1291.
Alpha tubulin:	-CTATTTATTTATCATGCGT	4	68	Sather '85 PNAS 82, 5695.
Beta tubulin:	-GTCCAAACGAATTATGCGC	4	67	''
T.congolense				
Trypanothione reductase:	-CAATCGCTTTTCTATGTCG	2,5,6	–	Shames '88 Biochem. 27, 5014.
Leishmania				
L.enriettii Membrane transport protein:	-ATTCACTAGAATCATGAGC	2,6	–	Cairns'89 PNAS 86, 7682.
Alpha-tubulin:	-TTCCTTGTCAACCATGCGT	2,4	57	Landfear '86 MBP 21, 235.
Beta-tubulin:	-CGGCTCTATCACGATGCGT	2,4	63	''
L.major Dihydrofolate reductase:	-GAGCACTACGAAGATGTCC	2,4,6	39	Beverley '86 PNAS 83, 2584.
L.tropica Dihydrofolate reductase:	-GAGCACTACGAAGATGTCC	2,3,6	43	Grumont '86 PNAS 83, 5387.
Crithidia				
C.fasciculata Dihydrofolate reductase:	-CTCGTTCTCAAAGATGTCA	2,4,6	59	Hughes '89 MBP 34, 155.
Giardia				
G.lamblia Beta-tubulin II:	TTAAAAATGCGT	4	100	Kirk-Mason '89 MBP 36, 87.

a: The sequences were shown as plus strand of DNA. When the cDNA was expected to possess a 5'leader longer than the determined sequence, the left side of the sequence in the appendix was marked by a hyphen.
b: Information used to identify the translational initiation site is given in the Start Data column where 1= Comparison of DNA sequence with amino acid sequence determined independently, 2 = Open reading frame analysis of genomic DNA, 3 = Open reading frame analysis of cDNA, 4 = 5' transcript mapping data and DNA sequence analysis, 5 = Analysis of *in vitro* transcription/transla tion products compared with DNA sequence, and 6 = Comparison of homologous genes.
c: Percent of the A- and T-residue contents of the 5' leader sequence.
d: Bibliographic data are given in the References column in condensed form: first name, year, journal, volume, and first page.
e: There is an intron between initiation site and second codon. Thus the separating intron was omitted and the actual translatable mRNA sequence is used.
f: The sequence is being deposited in the EMBL/GenBank data base (accession no. M57542).
*: Non-functional ATGs were found in the 5'-leader sequences.