214–221 Nucleic Acids Research, 1994, Vol. 22, No. 2

Characterization of transcription initiation, translation initiation, and poly(A) addition sites in the gene-sized macronuclear DNA molecules of *Euplotes*

Susmita Ghosh, John W.Jaraczewski[1], Lawrence A.Klobutcher* and Carolyn L.Jahn[1]
Department of Biochemistry, University of Connecticut Health Center, Farmington, CT 06030 and [1]Department of Cell, Molecular, and Structural Biology, Northwestern University Medical School, Chicago, IL 60611, USA

Received September 15, 1993; Revised and Accepted November 30, 1993 EMBL accession nos+

ABSTRACT

The DNA in the transcriptionally active macronucleus of the hypotrichous ciliate *Euplotes crassus* exists as short, linear molecules with each molecule encoding a single genetic function. Previous work has indicated that coding regions occupy the majority of macronuclear DNA molecules. In the present study we have defined the transcription initiation sites and poly(A) addition sites for a number of different macronuclear genes in *Euplotes crassus*. Our results indicate that mature transcripts represent all but ~100–200 bases of the non-telomeric sequences in macronuclear DNA molecules. We have also examined the sequences in the vicinity of transcription start sites, poly(A) addition sites, and translation initiation sites for *Euplotes* species genes in an attempt to define the *cis*-acting elements that control these processes. Our results indicate that some of the common sequence elements known to control these processes in higher eukaryotes are likely not utilized by *Euplotes* genes. The data do indicate the presence of other conserved sequences both preceding and at the site of poly(A) addition, as well as at the site of translation initiation. These conserved sequences may serve an analogous role in these organisms. Finally, we have found that most macronuclear DNA molecules have transcription initiation sites within 30 bp of the telomere, suggesting that the telomere may play a role in promoting transcription.

INTRODUCTION

Like other hypotrichous ciliated protozoa, *Euplotes crassus* has a macronucleus containing a large number of gene-sized DNA molecules (average size of ~2 kbp), each of which is present in multiple copies (reviewed in 1,2). All of these macronuclear DNA molecules have telomeres at their ends that consist of repeats of the octanucleotide 5′CCCCAAAA3′. Previous studies indicate that each macronuclear DNA molecule encodes a single genetic function. Numerous macronuclear DNA molecules containing genes of known function have now been analyzed and, in most cases, the coding regions occupy the vast majority of the macronuclear DNA molecule. The macronuclear DNA molecules of *Euplotes crassus*, the primary focus of the current study, appear to be particularly compact, with coding regions occupying all but ~200 bp of the macronuclear DNA molecules.

Despite the fact that macronuclear DNA molecules in this species have extremely short 5′ and 3′ non-translated sequences, little is known concerning the sequence elements involved in transcription initiation and RNA processing. To learn more about these signals, we have determined the length of the mRNAs for a number of genes of known and unknown function. We then carried out primer extension analyses to precisely define the transcription initiation sites of a number of genes, as well as determined a number of poly(A) addition sites. The current data, along with previously published results on other *Euplotes* genes, have been analyzed to determine potential transcription initiation, translation initiation, and polyadenylation control signals. Our results indicate that some of the common eukaryotic control signals are not used by *Euplotes*, but that other conserved sequences may be playing similar roles in the control of gene expression.

MATERIALS AND METHODS

RNA isolation

For the isolation of total RNA from vegetative *E.crassus* cells, the cells were grown to high density as described (3,4), but not allowed to starve. The cells were then filtered through a 100 μm Nitex membrane (Tetko, Inc., Elmsford, NY) and collected on a 15 μm Nitex to a final volume of 25 ml per liter of original culture. The cells were further concentrated by centrifugation at room temperature at 250×g for 1 minute. The supernatant

* To whom correspondence should be addressed

+ V01537–V01551 (incl.)

was removed by aspiration and the cells were either quick frozen for later processing or immediately lysed by the addition of 4M guanidinium isothiocyanate to a final volume of 10 ml per 0.5−1.0 ml of cell pellet. RNA was then isolated from this lysate either by centrifugation over a 5.7 M CsCl cushion (5) or by water saturated phenol/chloroform extraction (6). This crude RNA was either used directly for polyA+ RNA isolation, or further purified on a QUIAGEN column (Quiagen, Inc., Chatsworth, CA) following the manufacturer's protocol. Yield of total RNA was generally about 1 mg per liter of vegetative cells.

For polyA+ RNA isolation, oligo-dT cellulose (Pharmacia, Piscataway, NJ) chromatography was performed for two cycles as described by Sambrook *et al.* (7) with minor modifications. PolyA+ RNA yield was 7 μg per mg of total cell RNA.

Northern blot analyses

For Northern blot analysis, 50 μg of total cellular RNA was denatured by the glyoxal/DMSO method and electrophoresed on a 1.2% agarose gel prepared and run in 10 mM NaH_2PO_4, pH 7.0 (8). Blotting and deglyoxylation were done on Genescreen-Plus membrane (NEN Research Products, Boston, MA) according to the manufacturer's protocols. Prehybridization and hybridization of the blots was carried out at 65°C in 6× SSC (1× SSC = 0.15 M NaCl, 0.015 M sodium citrate, pH 7.0), 0.7% SDS, 1× Denhardt's solution (0.02% bovine serum albumin, 0.02% polyvinylpyrrolidone, 0.02% Ficoll), and 100 μg/ml salmon sperm DNA. The filters were washed as described previously (9).

Hybridization probes for Northern blots and other procedures were labeled with ^{32}P by the random hexamer priming method (10,11).

cDNA library construction and isolation of cDNA clones

The LEVR cDNA library was prepared from polyA+ RNA isolated from vegetatively growing *E.crassus* strain ST11 cells. 5 μg of polyA+ RNA was used to synthesize double-stranded directional cDNA using the UNIZAP-XR cDNA cloning kit (Stratagene, LaJolla, CA). This cDNA was then ligated to EcoRI + XhoI digested λ-UNIZAP-XR vector, and the ligated material was used to generate phage using the Gigapack Gold packaging system (Stratagene). 4.5×10^6 recombinant clones were generated, and 7×10^5 phage from the primary library were amplified to 4×10^{11} phage by growing them for 7 hrs in *E.coli* SURE™ cells (Stratagene) for use in further analyses.

The LEVR library was screened for cDNA clones corresponding to MACET-3, PGK, and V3 genes by the plaque lift hybridization method (7). cDNA clones of beta-tubulin, histone H4, actin, and RPL29 were obtained by amplification of cDNAs by the RACE procedure essentially as described by Frohman *et al.* (12). The 'polyT adapter primer' and 'upstream adapter primer', synthesized by Operon Technologies Inc. (Alameda, CA), were identical in sequence to those described by Frohman *et al.* (12). The initial cDNA reaction was performed using the Superscript Kit (BRL Inc., Gaithersburg, MD) with the polyT adapter primer and 1 μg of total RNA. For the PCR reactions, 2.5% of the cDNA reaction was amplified using 25 pmol each of the upstream adapter primer and a 5′ gene-specific primer. The 5′ gene-specific primers were:

Beta-tubulin	5′GCTTCTACCTTCATCGG3′
Actin	5′GGAGAGGCTCTACAAGG3′
Histone H4	5′GCCAAGAGACACGCCAAGAAG3′
RPL29	5′AAGCACCACCACAGAATTAATGTG3′

Thirty-five cycles of PCR amplification (13) were performed using Ampli-Wax beads and the 'hot-start' technique (Perkin-Elmer Cetus, Norwalk, CT). The final 100 μl reaction contained 10 mM Tris−HCl (pH 8.3), 1.5 mM $MgCl_2$, 50 mM KCl, 200 μM dNTPs, 0.1 mg/ml gelatin, and the primers. The first PCR cycle was done at a denaturing temperature of 95°C for 90 sec, and all subsequent denaturing steps were at 94°C for 30 sec. Annealing (20 sec per cycle) was carried out at a temperature appropriate for the primer with the lower T_m (calculated by subtracting 6°C from the T_m, where $T_m = 4(G+C)+2(A+T)$). Extensions were carried out at 72°C for 45 sec per cycle. The PCR products were then cloned into the Sma I site of the pKS+ vector (Stratagene) after purification from an agarose gel using USBIOCLEAN (United States Biochemical Corp., Cleveland, OH).

For the isolation of the random cDNA clones an aliquot of the LEVR phage library was converted to a plasmid library (14). This was done by first infecting *E.coli* Bluescript SK-XL-1 Blue cells in the presence of ExAssist™ helper phage (Stratagene) to allow in vivo excision and packaging of the phagemid portion of the vector. The phage produced were then used to infect *E.coli* SOLR™ cells (Stratagene) to generate bacteria carrying plasmids with cDNA inserts, from which the random clones (pEVR clones) were selected. All of the above steps were performed according to the manufacturer's protocol with some minor modifications.

Primer extension analysis

The following oligonucleotides were synthesized on a Cyclone dual column DNA synthesizer (Biosearch, Inc.) for use in primer extension analyses of the indicated genes:

Actin	5′ATTGTCGGTGGAGCTGGA3′
Histone H4	5′GGTCGGAGCCAAGAGACACG3′
MACET-3	5′GACGGGGTCGTTACTGATGCTTTAGGAAAC3′
PGK	5′ATGATTAAGAACAAGAGAGTCTTAGTCAGAGTA3′
RPL29	5′CACATTAATTCTGTGGTGGTGCTTTGGACCAGCC3′
Beta-tubulin	5′GTTGATTCTTTCGAGCT3′

Prior to use in primer extension, all oligonucleotides were run on 15% urea-acrylamide gels and purified as described (7)

Primer extension analyses were performed using modifications of published protocols (7,8). $1-5\times10^5$ cpm of the 5′end-labeled oligonucleotide primer was annealed overnight with 25 μg of total RNA, 2 μg of polyA+ RNA, or 20 μg of tRNA (control) in the presence of 400 mM NaCl, 40 mM PIPES (pH 6.8) and a varying concentration of formamide depending upon the length of the primers (see below). The annealed mixture was then ethanol precipitated and dissolved in 20−25 μl of reverse transcriptase reaction mixture consisting of 2.5 μg/ml actinomycin D, 0.25 mM of each nucleotide triphosphate, 25 U/ml RNasin (Promega Corp., Madison, WI), 10 U of MMLV reverse transcriptase (United States Biochemical Corp.), 50 mM Tris−HCl (pH 8.3), 40 mM KCl, 6 mM $MgCl_2$, and 10 mM DTT. The reverse transcription reaction was carried out for 1 hr. RNase A and salmon sperm DNA were then added to concentrations of 20 μg/ml and 10 μg/ml, respectively, and the sample was incubated a further 15 min at 37°C;. The reactions were phenol/chloroform extracted, ethanol precipitated, and redissolved in sequencing gel loading buffer (7). The

oligonucleotide annealing conditions and reverse transcription reaction temperatures were adjusted based on the lengths of the oligonucleotides used. The concentrations of formamide used in oligonucleotide annealing, the temperatures of the annealing reaction, and the temperature for the reverse transcriptase reactions, respectively, for each of the genes analyzed were as follows: actin, 50%, 45°C, and 42°C; histone H4, 40%, 37°C, and 37°C; MACET-3, 50%, 45°C, and 42°C; PGK and RPL29, 80%, 45°C, and 42°C. The beta-tubulin primer extension analysis used an alternate annealing procedure without formamide as described for primer extension RNA sequencing by Bektesh *et al.* (15).

The primer extension products were analyzed on 6% Hydrolink (J.T. Baker, Inc., Phillipsburg, NJ) sequencing gels. Size standards consisted of DNA sequence ladders generated by sequencing a clone of the corresponding macronuclear DNA molecule, or an in vitro generated transcript from a macronuclear clone, using the same oligonucleotide primer employed in the primer extension analysis.

DNA sequencing and DNA sequence analysis

Plasmids DNA preparations used in sequencing were made using the Magic™ Miniprep DNA Purification System (Promega, Madison, WI). Dideoxy sequencing reactions were performed using a Sequenase version 2.0 kit (United States Biochemical Corp.) and the sequencing reactions were analyzed on 6% Hydrolink gels (J.T.Baker, Inc.,).

To identify conserved sequences near poly(A) addition sites and translation initiation sites we used a procedure similar to that described by Goodrich *et al.* (16) that compensates for the overall base composition of the region. The frequency of the individual bases, as well as purines, pyrimidines, was first calculated for all sequences in the region under analysis. The mean expected frequency of each base at a given position was then calculated as nf_b, where n is the number of different sequences being analyzed and f_b is the frequency of the base (A,C,G,T,R, or Y). For a random sequence, the expected frequency of a base at a given position was assumed to follow a binomial distribution, so that the expected standard deviation could be calculated as the square-root of nf_bf_x, where f f_b again represents the overall frequency of a given base (e.g., A) and f_x represents the combined overall frequency of the remaining bases (e.g., T,C, and G). Positions in the sequence where the observed frequency of a base exceeded the mean plus twice the standard deviation were then considered significant.

The following conventions were employed in developing consensus sequences. Highly conserved residues are defined as those that were both statistically significant by the above criteria and were present in greater than 65% of the sample. They are denoted by upper case letters. Lower case letters in the consensus repreent positions that were statistically significant but the base was not present in greater than 65% of the sequences in the sample. In addition, the consensus includes positions that were not themselves statistically significant, but met the following two criteria: 1) they were immediately adjacent to statistically significant positions, and 2) a particular base was present in greater than 65% of the sample. These positions are also shown in lower case in the consensus.

GenBank (release 71.0) searches with the putative polypeptides encoded by the random cDNA clones were performed using MacVector version 3.5 software (International Biotechnologies, Inc., New Haven, CT) and the pam250 scoring matrix. Only those matches that generated scores at least ten standard deviations above the mean were considered significant.

RESULTS AND DISCUSSION

Sizes of mRNAs from macronuclear DNA molecules

Numerous studies have indicated that most macronuclear DNA molecules possess coding regions that occupy the majority of their length (Table I). We wished to extend this type of analysis by surveying the sizes of transcripts produced from different *E.crassus* macronuclear DNA molecules. Previous studies had determined the sizes of mRNAs from the *E.crassus* macronuclear DNA molecules encoding actin and beta-tubulin (17,18) and found the transcripts to be about 100 b smaller than their respective macronuclear DNA molecules (Table I). We carried out additional Northern hybridization analyses using cloned macronuclear DNA molecules encoding histone H4 (18), a thiol reductase homolog (MACET-3; 19), phosphoglycerate kinase (PGK; R.Pearlman and L.Klobutcher, unpublished results), and a protein kinase homolog (V3; 9, Hale and Klobutcher, unpublished results) as probes (Table I; data not shown).

Most of the transcripts detected for the genes of known function were of a similar size to the macronuclear DNA molecules encoding them (Table I). With one exception, the difference in the length of the transcript and the corresponding macronuclear DNA molecule did not exceed 200 b. The single observed exception to this general organization of macronuclear DNA molecules is the histone H4 gene. The histone H4 macronuclear DNA molecule was previously shown to contain a large 5′-untranslated region of about 1.3 kbp (18). Our results indicate that most of this 5′ region also does not appear in the mature message, but may be involved in the transcriptional control of this gene during the cell cycle as suggested previously (17,20).

Overall, the results indicate that the vast majority of sequence information present in a macronuclear DNA molecule is present in its transcript. Allowing for both a poly(A) tail on the transcript and telomeres on the macronuclear DNA molecule, the mature transcripts contain all but ~100–200 bp of the non-telomeric DNA in a macronuclear DNA molecule. These short stretches of sequence that do not appear in the mature transcripts would presumably include the majority of transcriptional control sequences, and possibly origins of DNA replication and the sequences that specify chromosome fragmentation during the formation of the macronucleus.

Table I. Sizes of macronuclear DNA molecules, coding regions, and transcripts

Gene	Macronuclear DNA Size (kbp)[1]	Size of Coding Region (kbp)	Transcript Size (kb)
Actin	1.303	1.140	1.2
Histone H4	1.874	0.320	0.46
β-Tubulin	1.521	1.337	1.5
PGK	1.416	1.214	1.3
MACET-3	0.604	0.420	0.50
V3	1.807	1.413	1.6
V4[2]	1.6	nd	1.6

nd = not determined

[1]Sizes of macronuclear DNA molecules include the double stranded regions of the telomeres, but not the single-stranded tails.

[2]The 1.6 kbp V4 macronuclear DNA molecule (9) and its transcript cross-hybridize with the V3 hybridization probe.

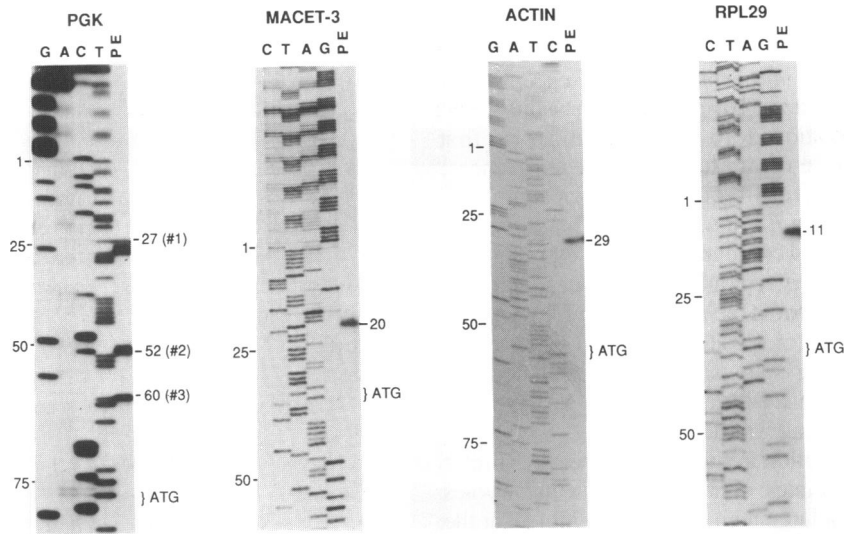

Figure 1. Primer extension analysis on the transcripts of the PGK, MACET-3, actin, and RPL29 genes. Primer extension reactions are shown (lanes labeled PE) along with a DNA sequencing ladder that served as the size standards. The DNA sequencing ladder was generated by sequencing a clone of the corresponding macronuclear DNA molecule, or for the PGK gene an in vitro generated transcript derived from a macronuclear clone, using the same oligonucleotide primer used in the primer extension analysis. Numbering of the sequences begins with the first base following the telomeric repeats. The positions of transcription initiation sites are indicated along with the position of the translation initiation codon (ATG) of each gene.

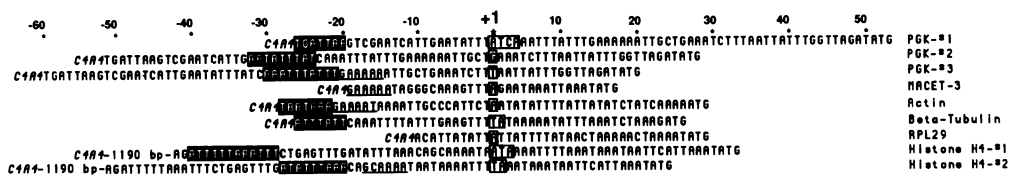

Figure 2. Transcription start sites of *E.crassus* genes. The sequences in the vicinity of transcription initiation sites (boxed) are shown, from the C_4A_4 telomeric repeats through the ATG initiation codon. Numbering of the sequences in this case begins with the first transcribed base in an initiation region; upstream bases are negatively numbered. Genes with multiple regions of transcription initiation (PGK and histone H4) are shown multiple times, so as to align the transcription initiation regions. AT-rich sequences that may function as TATA boxes are shown with a black background and close matches to the sequence 5'GAAAAA3', which may function as a promoter element, are underlined.

Transcription initiation sites

To more precisely define the transcription units of macronuclear DNA molecules, and possibly define promoter elements, we used primer extension analysis to determine the transcription start sites of the *E.crassus* PGK, MACET-3, ribosomal protein L29 (RPL29; D.Shippen and C.Jahn, manuscript in preparation), and actin genes (Fig. 1 and 2), as well as the histone H4 and beta-tubulin genes (data not shown). Single transcription initiation sites were observed 20, 29, and 11 bp from the telomeric repeats for the MACET-3, actin, and RPL29 genes, respectively. The beta-tubulin gene displayed a pair of adjacent start sites 27 and 28 bp from the telomere. In contrast, the PGK and histone H4 genes displayed multiple regions of transcription initiation. The PGK gene has a cluster of 4 sites beginning 27 bp from the telomere, as well as strong start sites at both 52 and 60 bp from the telomere. Histone H4 displayed two regions of initiation; three sites beginning 1233 bp from the telomere, and two adjacent sites 1242 bp from the telomere.

The regions upstream of the transcription initiation sites were searched for known common transcriptional control sequences or other possible shared sequence elements. Strong matches to the 'TATA box' sequence (5'TATA(A/T)A3'), which normally begins 25−35 bp upstream of the transcription initiation site, are not observed at the appropriate position in any of the *E.crassus* genes. These results are similar to other studies of ciliate genes and it has been suggested that a TATA box may not be an essential promoter element in these organisms (21,22). Based on the current data it is difficult to completely rule out a role for the TATA box, as many of the genes do have AT-rich sequences beginning about 30 bp from their start sites (Fig. 2, black-boxed sequences) that may be serving the function of the TATA box. It is quite clear from the analyses of two of the genes, however, that an appropriately positioned TATA box is not an absolute requirement for transcription in all *Euplotes* genes. Both the MACET-3 and RPL29 genes have their telomeric repeats positioned within 20 bp of their start sites, precluding a properly positioned TATA box. The extremely short upstream regions of these two genes suggests that they may possess transcriptional control elements in the vicinity of the transcription initiation site itself, similar to some of the TATA-less promoters of genes in higher organisms (reviewed in 23). There is, however, no obvious sequence similarity to any of the promoter elements of this class that have been defined to date.

The upstream regions have also been examined for matches to the sequence 5'TATCCAATCARA3', which has been noted

upstream of a number of *Tetrahymena* genes (mainly histone genes) and other ciliate genes (22). No close matches to this sequence are found in any of the genes, with the exception of the histone H4 gene. The histone H4 gene has the sequence 5′ GAGCCAATCAGA3′ positioned 76 bp upstream of its first transcription initiation site. In regard to other shared sequences, we have been unable to define any sequence block of significant length that is shared by all the *E.crassus* genes. The search for common sequence elements is confounded by the extreme AT-richness of the non-transcribed regions. The best candidate for a transcriptional control element is the sequence 5′GAAAAA3′. Close matches to this sequence are positioned 17−21 bp upstream of four of the transcription start sites (Fig. 2, underlined sequences). Additional genes will have to be analyzed to determine if this is a recurring motif.

The actual sites of transcription initiation are generally similar to other eukaryotes and ciliates. Transcription in eukaryotes usually initiates at an A residue, and a tendency to initiate at the A residues in the sequences 5′TA3′ or 5′YAA3′ has been noted for ciliates (21,24). Many of the *E.crassus* transcripts start at an A residue that is preceded by a T, but a number of sites do not follow this convention (Fig. 2).

Perhaps the most surprising aspect of the results is the fact that all of the genes examined, with the exception of the histone H4 gene, have a transcription initiation site positioned within 29 bp of the telomeric repeat sequences. This raises the interesting possibility that the telomeric repeat sequences, and perhaps the proteins associated with them (reviewed in 25), may be playing a role in either defining the start of transcription or enhancing transcription initiation. This is surprising as studies on yeast indicate that placing a gene within a telomeric context represses transcriptional activity (e.g. 26,27). This type of telomeric repression would of course be lethal in hypotrichs, where all of the functional genes are, in effect, in a telomeric context. Perhaps the difference in the two organisms is that the yeast telomeres consist not only of terminal simple repeat sequences similar to those in *Euplotes*, but also of larger subterminal repetitive elements (28). It may be that the latter sequence elements are required for transcriptional silencing.

Translation initiation sites

The mapping of transcription start sites also allowed us to determine the length of the 5′-non-translated regions of mRNAs in *E.crassus*. The 5′-non-translated regions are extremely short, ranging from 14−48 b in length (Fig. 2). The sizes of these leader regions fall at the low end of the spectrum observed for vertebrate RNAs (29). They do all exceed 10 b, however, which appears to be the lower limit for efficient translation in vertebrates (29).

We then went on to determine if a conserved sequence is present in the vicinity of the translation initiation site as has been observed for a number of organisms (reviewed in 29,30). Figure 3A displays the sequences in the vicinity of the translation initiation codon for the sixteen *E.crassus* genes and genes of other *Euplotes* species that have been reported in the literature or that have recently been determined. In order to identify potential conserved sequences at or near the site of translation initiation we carried out a statistical analysis to search for non-random base composition on the 20 positions preceding through the 10 bp following the ATG codon. The method employed was similar to that described by Goodrich *et al.* (16), in that it looks for statistically significant deviations from random base composition at positions a given distance from a landmark, while taking into

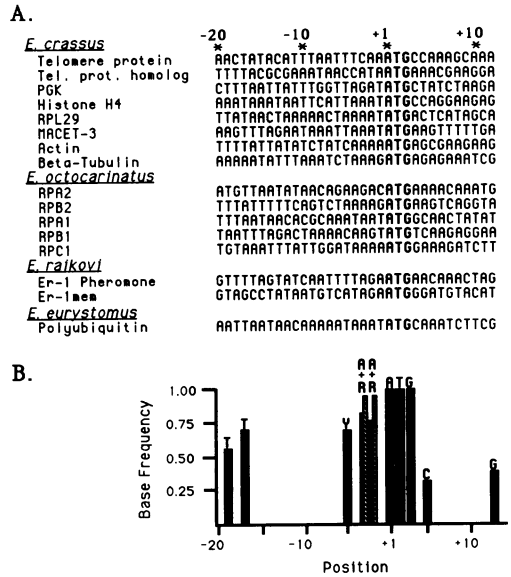

Figure 3. A.) The sequences in the vicinity of the translation initiation codons of 16 *Euplotes* species genes are shown with the sequences aligned at the ATG initiation codon (bold type). The abbreviations and sources of sequence information for the genes not previously noted in the text are: telomere protein and telomere protein homolog (31); RPA1, RPB1, RPC1, RPA2, and RPB2 are the largest subunits of RNA polymerase I, II, and III, and the second largest subunits of RNA polymerase I and II, respectively (32,33); Er-1 pheromone (34); Er-1mem, membrane bound form of Er-1 pheromone (35), and; polyubiquitin (36). **B.)** Plot of base frequency versus the positions in sequences surrounding the translation initiation codon. The region analyzed was the 20 bp preceeding the ATG codon, through the 10 bp following the ATG codon. Only those positions where statistically significant deviations from random base composition were observed are plotted. Significant positions for A, C, G, or T are indicated by solid black bars and significant positions for purines or pyrimidines are indicated by hatched bars. Note that a single position may show significant base composition for more than one category. The base composition of the region anlayzed was: G = 0.14, A = 0.45, T = 0.31, and C = 0.10. Considering this base composition, and a sample size of 16, a base must occur at a given position in excess of the following values to be considered statistically significant (see Materials and Methods): G = 5.1, A = 11.2, T = 8.6, C = 4.0, Y = 10.4, and U = 13.4.

account the general base composition of the region under analysis (see Materials and Methods). In this case the ATG codon that initiates translation constitutes the landmark. Positions where the frequency of a given base (or purines and pyrimidines) exceeded the expected mean plus two standard deviations were considered significant.

Positions where a significant deviation from random base composition was observed are displayed graphically in figure 3B. In carrying out this type of statistical analysis on a large number of positions, we expect to observe a number of positions that exceed the significance cut-off by chance alone. Some of the isolated significant positions observed in our analysis (e.g., positions −19 or −17) may be the result of such chance deviations. However, a cluster of sites of nonrandom base composition immediately precedes the initiation codon, and suggests that these positions are truly conserved. Inspection of the sequence of this region suggests a consensus sequence of $5'Y_{69}a_{69}A_{81}A_{75}N\underline{ATG}3'$ for *Euplotes* initiation sites (subscripts indicate the percentage of sequences in the sample that conform to the consensus; see Materials and Methods for the conventions used in developing consensus sequences).

The *Euplotes* consensus sequence is quite different from the consensus sequence 5′GCCGCC(A/G)CC$\underline{ATG}$G3′ that has been

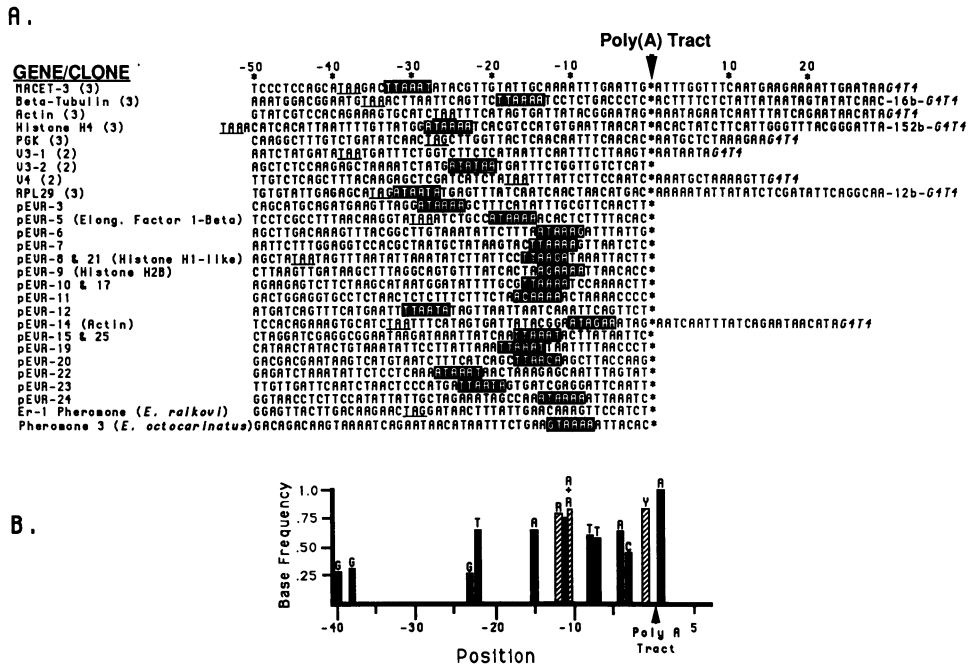

Figure 4. A.) Sequences surrounding poly(A) addition sites for *Euplotes* genes. The sequences are aligned based on the position of the poly(A) tract observed in the cDNA clones. For the cases where the sequence of the corresponding macronuclear DNA molecule is known, both the sequence preceding and following the position of the poly(A) tract in the cDNA clone is shown, and the number of cDNA clones analyzed for each gene is shown in parentheses. Translation termination codons are underlined and G_4T_4 telomeric repeats are indicated in italics. Perfect matches, and 5 of 6 bp matches, to the proposed upstream sequence element 5′(A/T)TAAAA3′ are shown with a black background. For the random cDNA clones (pEVR clones), putative coding functions revealed by GenBank computer searches are indicated in parentheses. **B.)** Plot of the base frequency versus positions in sequences surrounding the poly(A) tract. Only those positions where statistically significant deviations from random base composition were observed are plotted. Significant positions for A, C, G, or T are indicated by solid black bars and significant positions for purines or pyrimidines are indicated by hatched bars. Note that a single position may show significant base composition for more than one category. The region anlayzed was from −40 to +7, which has the following base composition: G = 0.12, A = 0.38, T = 0.35, and C = 0.16. For the 40 bases preceeding the poly(A) addition site, where n = 27, a base must occur at a given position in excess of the following values to be considered statistically significant (see Materials and Methods): G = 6.4, A = 15.3, T = 14.3, C = 8.1, Y = 20.9, and U = 20.6. For the 7 bases downstream of the poly(A) addition site, where n = 9, a base must occur at a given position in excess of the following values to be considered statistically significant: G = 2.9, A = 6.3, T = 6.0, C = 3.6, Y = 7.5, and U = 7.4.

deduced for vertebrate translation initiation sites (29,30). However, the proposed *Euplotes* consensus shares an A residue at position −3 with the vertebrate sequence, and this position has been shown to play a major role in defining initiation sites in vertebrates (reviewed in 30). Although the consensus sequence we have derived differs from that of vertebrates, it is similar to the proposed consensus sequences for a number of lower eukaryotic organisms, which tend to have a number of A residues preceding the initiation codon (37). The *Euplotes* consensus is especially similar to the consensus sequences 5′AAAATG3′ proposed for protozoa in general (37), and 5′AAAATGG3′ or 5′UY$(A)_{3-4}$ATGG3′ proposed for the ciliates (21,22). Our results are thus consistent with the notion that lower eukaryotes do not conform to the vertebrate translation initiation consensus sequence.

Poly(A) addition sites

To analyze poly(A) addition sites, we isolated and sequenced cDNA clones corresponding to seven macronuclear DNA molecules whose entire sequences had been determined. In most cases, three independent cDNA clones were analyzed for each macronuclear DNA molecule. In addition, we isolated and sequenced the 3′ ends of 19 clones chosen randomly from an *E.crassus* vegetative cDNA library (pEVR clones). The sequences of the *E.crassus* genes in the vicinity of the poly(A) addition site are shown in figure 4A, along with the previously determined sequences of cDNA clones for the *E.raikovi* Er-1 pheromone gene (34) and the *E.octocarinatus* pheromone 3 gene (38), which are the only other poly(A) addition sites determined in the *Euplotes* species group.

Examination of the sequences revealed a number of features of poly(A) addition in *E.crassus*. First, the data suggest that single poly(A) addition sites are used for most genes, as the poly(A) addition site was at the same position in each of the cDNA clones for MACET-3, β-tubulin, histone H4, PGK, and RPL29 (Fig. 4A). In addition, in three instances, two independent random cDNA clones were derived from the same mRNA (Fig. 4A), and in each case both clones used the same poly(A) addition site. Two different sites of poly(A) addition were identified for actin. Three actin cDNA clones produced from RNA from cells undergoing macronuclear development by the RACE PCR procedure (12) all showed the same site of poly(A) addition. However, one of the random cDNA clones (pEVR-14) isolated from the vegetative cell cDNA library also was derived from the actin gene and employs a poly(A) addition site 6 b downstream from the other three clones. It is currently unclear whether these two poly(A) addition sites are used in a developmental-specific manner or if both sites are utilized at each developmental stage. The V3 macronuclear DNA molecule also presented a more complex situation. In screening the vegetative cell cDNA library

with the V3 gene we isolated not only cDNA clones corresponding to the previously sequenced macronuclear DNA molecule (allele V3-1; 9), but also clones corresponding to a second V3 allele that we have subsequently identified (allele V3-2; 39, Bernaski, Hoppe, and Klobutcher, unpublished results) and cDNA clones corresponding to the 1.6 kbp macronuclear DNA molecule V4 (9). The 1.6 kbp V4 macronuclear DNA molecule, based on partial DNA sequence analysis, appears to encode a protein kinase that is 89% identical to the V3 protein kinase at the amino acid level (Hale and Klobutcher, unpublished results). Our results suggest that the two alleles of V3 utilize different poly(A) addition sites, which may be related to base substitutions present in the 3′ non-coding regions of the two alleles (Fig. 4A).

For the genes of known function, the poly(A) tracts in the cDNA clones are all located within 51 bp of the termination codon. The regions 3′ of the poly(A) addition sites are also generally short. With the exception of the histone H4 gene, which has 181 bp of sub-telomeric sequence after the poly(A) addition site, none of the regions exceed 45 bp (Fig. 4A). Allele 1 of the V3 gene has only 7 bp of sub-telomeric DNA following the poly(A) site, the shortest such region observed to date for a hypotrich gene.

A number of sequence elements have been found to control poly(A) addition in higher eukaryotes (reviewed in 40) and we examined the *Euplotes* sequences to determine if similar motifs were present. Poly(A) addition tends to occur at a 5′CA3′ dinucleotide in higher eukaryotes. Based on the *Euplotes* genes of known function, the poly(A) tail can always be viewed as being added prior to, or following, an A residue in the primary transcript (Fig. 4A), which is similar to the situation for animal genes. To determine if there was further sequence conservation at the site of poly(A) addition, or at other positions relative to the poly(A) tracts, we again carried out a statistical analysis of base composition (Figure 4B) as was done for translation initiation sites. The interval analyzed included the 40 bp preceding the position of the poly(A) tract, plus 7 bp downstream. The most striking feature in this analysis is a cluster of statistically significant positions surrounding the site of the poly(A) tract in the cDNA clones. The data suggest a consensus sequence of $5'a_{63}c_{44}NY_{81}A_{100}$-(poly(A) add. site)-$a_{67}3'$. This sequence is, in effect, an extended version of the 5′CA3′ sequence often present at the site of polyadenylation in vertebrate transcripts. We suggest in our consensus that transcript cleavage and poly(A) addition actually occur after the universally conserved A residue, which is consistent with the limited data available from in vitro studies of animal systems (41,42).

A second sequence element that is involved in directing poly(A) addition in higher eukaryotes is the sequence 5′AATAAA3′, usually located 10−30 bp upstream of the poly(A) addition site. Perfect matches to this sequence motif are found within the 50 bp upstream of the poly(A) addition site in only 3 (pEVR-6, pEVR-22, and pEVR-24) of the 27 cases we have examined (Fig. 4A). As essentially all base changes to this sequence element in higher eukaryotes greatly reduce the efficiency of poly(A) addition (43), the current results suggest either that this sequence element is not used in *Euplotes* or that the sequence requirement is relaxed. The absence of the 5′AATAAA3′ upstream element in *Euplotes* would not be surprising, as this motif does not appear to be required for poly(A) addition in both yeast and plants (reviewed in 40). In addition, the sequence 5′TAAAC3′, which has been suggested to serve as an upstream poly(A) addition signal in the hypotrich *Stylonychia lemnae* based on a limited data set (44), is generally not found upstream of the *Euplotes* poly(A) addition sites.

The AT-richness of the *Euplotes* sequences makes it difficult to identify an alternate conserved sequence that can reside at a variable distance upstream of the poly(A) additions sites. The most likely candidate that we have been able to discern is the sequence 5′(A/T)TAAAA3′. Our reasons for suggesting this sequence are twofold. First, in the statistical analysis of the poly(A) addition site sequences, three positions between −11 and −15 displayed non-random base composition for A or purine residues (Fig. 4B). Inspection of the individual sequences in this region indicates that a purine-rich (primarily A residues) stretch of bases often exists in the vicinity of these positions. Second, of the 27 poly(A) addition sites, 22 have either an exact match or a 5 of 6 position match to this sequence located 10−33 bases upstream of the poly(A) addition site (Fig. 4A). The remaining sequences have related, but more diverged, versions of the 5′(A/T)TAAAA3′ element. The proposed sequence element is, however, AT-rich itself, so its frequent appearance in the upstream region may be coincidental. Functional studies will clearly be required to substantiate the involvement of this sequence.

Finally, poly(A) addition in higher organisms also appears to require a downstream element that is often a GT-rich sequence (reviewed in 40). Inspection of the sub-telomeric sequences downstream of poly(A) addition sites in *Euplotes* shows that obvious GT-rich sequences are usually not present (Fig. 4A). The telomeric repeats themselves are, however, solely composed of G and T residues on the non-template strand. That is, the G_4T_4 repeats would be part of the primary transcript if transcription does not actually terminate for hypotrich genes, but, instead, the polymerase simply proceeds to the end of the macronuclear DNA molecule and 'falls off'. These telomeric repeats on the primary transcript might then function as the downstream sequence element for polyadenylation. We are currently determining if transcription does indeed proceed into the telomeric repeats at the ends of macronuclear DNA molecules.

Putative coding functions of the random cDNA clones

In the sequence analysis of the 3′ ends of the random cDNA clones, 150−250 b of sequence information was generally obtained from each clone (the complete sequences have been deposited in GenBank under accession numbers U01537 through U01551). These regions were translated in all three possible reading frames, and the peptide encoded by the longest open frame for each clone was used to search the GenBank database. Four of the nineteen random cDNA clones displayed significant homology (i.e., matching scores at least 10 standard deviations above the mean) to proteins in the data base. This includes clone pEVR-14, which, as noted above, encodes actin. In addition, a carboxy terminal peptide of 54 amino acids predicted from the sequence of clone pEVR-5 shared 56% sequence identity with the carboxy terminal region of the elongation factor-1-beta protein of *Artemia salina* (45), and a putative 50 amino acid carboxy terminal peptide encoded by clone pEVR-9 showed ~80% sequence identity with the carboxy termini of numerous histone H2B proteins, including those of the ciliate *Tetrahymena thermophila* (46). A 47 amino acid peptide predicted from clone pEVR-8 shared significant homology with sea urchin late histone H1 proteins (e.g. 47). In this latter case it is unclear as to whether pEVR-8 actually encodes the *E.crassus* equivalent of a histone

H1, because the sequence identity was only 35% and the homology was primarily limited to lysine residues.

CONCLUSION

It has been common practice in previous studies reporting the sequence of a hypotrich macronuclear DNA molecule, usually without information on the mRNA, to note sequence elements known to control transcription or RNA processing in other organisms. Our results provide a strong indication that this is not justified. The sequences upstream of transcription initiation sites of *E.crassus* genes do not appear to contain conventional eukaryotic transcriptional control elements. Indeed, the extremely short upstream regions suggest that transcriptional control of macronuclear genes may be quite unusual and possibly involve the telomere. Similarly, the cis acting sequences controlling poly(A) addition appear to differ from those in higher eukaryotes.

Our results do indicate that there are conserved sequences both in the vicinity of translation initiation sites and poly(A) addition sites. However, there are a number of potential limitations to our analysis. First, our sample sizes are relatively moderate, and as more genes are examined it may be necessary to modify or reevaluate the significance of our proposed consensus sequences. Second, the samples are probably biased towards the inclusion of highly expressed genes. These genes may be atypical of the general population. Third, the non-coding regions of macronuclear DNA molecules also probably contain sequence elements involved in the chromosome fragmentation process of macronuclear development and macronuclear origins of DNA replication. It is possible that some of the conserved sequence elements we have defined are actually involved in these processes rather than translation initiation or poly(A) addition. We consider this unlikely, as the consensus sequences that have been derived are at, or near, their proposed functional sites.

Ultimately, the significance of the proposed sequence elements will need to be verified by mutational analyses. This could be achieved by either developing in vitro systems or by cell transformation. A reliable transformation system has not yet been developed for a hypotrichous ciliate, but recent improvements in the ease and efficiency of transformation of *Tetrahymena* (48) may now make it feasible to introduce exogenous DNA into *Euplotes*. The data we have obtained should aid in the rational design of vectors to achieve this goal.

ACKNOWLEDGEMENTS

We thank Ms Mary Ellen Jacobs for her critical reading of the manuscript and Mr Ken Hoppe for expert technical assistance. This work was supported by Public Health Service grants GM33277 to L.A.K. and GM37661 to C.L.J.

REFERENCES

1. Klobutcher,L.A. and Prescott,D.M. (1986) In Gall, J. (ed.), The Molecular Biology of Ciliated Protozoa. Academic Press, New York, pp. 111–154.
2. Prescott, D.M. (1992) *BioEssays*, 14, 317–324.
3. Roth, M., Lin, M. and Prescott, D.M. (1985) *J. Cell Biol.*, 101, 79–84.
4. Klobutcher, L.A., Turner, L.R. and LaPlante, J. (1993) *Genes & Dev.*, 7, 84–94.
5. Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) *Biochemistry*, 18, 5294–5299.
6. Chomczynski, P. and Sacchi, N. (1987) *Anal. Biochem.*, 162, 156–159.
7. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
8. Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (1993) Current Protocols in Molecular Biology. Current Protocols, USA.
9. Baird, S.E., Fino, G.M., Tausta, S.L. and Klobutcher, L.A. (1989) *Mol. Cell. Biol.*, 9, 3793–3807.
10. Feinberg, A.P. and Vogelstein, B. (1983) *Anal. Biochem.*, 132, 6–13.
11. Feinberg, A.P. and Vogelstein, B. (1984) *Anal. Biochem.*, 137, 266–267.
12. Frohman, M.A., Dush, M.K. and Martin, G.R. (1988) *Proc. Natl. Acad. Sci.* USA, 85, 8998–9003.
13. Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N. (1985) *Science*, 230, 1350–1354.
14. Short, J.M., Fernandez, J.M., Sorge, J.A. and Huse, W.D. (1988) *Nucleic Acids Res.*, 16, 7583–7600.
15. Bektesh, S., VanDoren, K. and Hirsh, D. (1988) *Genes & Dev.*, 2, 1277–1283.
16. Goodrich, J.A., Schwartz, M.L. and McClure, W.R. (1990) *Nucleic Acids Res.*, 18, 4993–5000.
17. Harper, D.S. and Jahn, C.L. (1989) *Gene*, 75, 93–107.
18. Harper, D.S. and Jahn, C.L. (1989) *Proc. Natl. Acad. Sci. USA*, 86, 3252–3256.
19. Klobutcher, L.A., Turner, L.R. and Peralta, M.E. (1991) *J. Protozool.* 38, 425–427.
20. Herrick, G. (1992) *J. Protozool.*, 39, 309–312.
21. Horowitz, S., Bowen, J.K., Bannon, G.A. and Gorovsky, M.A. (1987) *Nucleic Acids Res.*, 15, 141–160.
22. Brunk, C.F. and Sadler, L.A. (1990) *Nucleic Acids Res.*, 18, 323–329.
23. Weis, L. and Reinberg, D. (1992) *FASEB J.*, 6, 3300–3309.
24. Williams, K.R. and Herrick, G. (1991) *Nucleic Acids Res.*, 19, 4717–4724.
25. Henderson, E.R. and Larson, D.D. (1991) *Curr. Opinion Genet. and Dev.*, 1, 538–543.
26. Gottschling, D.E., Aparicio, O.M., Billington, B.L. and Zakian, V.A. (1990) *Cell*, 63, 751–762.
27. Renauld, H., Aparicio, O.M., Zierath, P.D., Billington, B.L., Chhablani, S.K. and Gottschling, D.E. (1993) *Genes & Dev.*, 7, 1133–1145.
28. Chan, C.S.M. and Tye, B.-K. (1983) *Cell*, 33, 563–573.
29. Kozak, M. (1987) *Nucleic Acids Res.*, 15, 8125–8132.
30. Kozak, M. (1989) *J. Cell Biol.*, 108, 229–241.
31. Wang, W., Skopp, R., Scofield, M. and Price, C. (1992) *Nucleic Acids Res.*, 20, 6621–6629.
32. Kaufmann, J. and Klein, A. (1992) *Nucleic Acids Res.*, 20, 4445–4450.
33. Kaufmann, J., Florian, V. and Klein, A. (1992) *Nucleic Acids Res.*, 20, 5985–5989.
34. Miceli, C., La Terza, A. and Melli, M. (1989) *Proc. Natl. Acad. Sci. USA*, 86, 3016–3020.
35. Miceli, C., La Terza, A., Bradshaw, R.A. and Luporini, P. (1992) *Proc. Natl. Acad. Sci. USA*, 89, 1988–1992.
36. Hauser, L.J., Roberson, A.E. and Olins, D.E. (1991) *Chromosoma*, 100, 386–394.
37. Cavener, D.R. and Ray, S.C. (1991) *Nucleic Acids Res.*, 19, 3185–3192.
38. Meyer, F., Schmidt, H.J., Plumper, E., Hasilik, A., Mersmann, G., Meyer, H.E., Engstrom, A. and Heckmann, K. (1991) *Proc. Natl. Acad. Sci. USA*, 88, 3758–3761.
39. Tausta, S.L., Turner, L.R., Buckley, L.K. and Klobutcher, L.A. (1991) *Nucleic Acids Res.*, 19, 3229–3236.
40. Wahle, E. and Keller, W. (1992) *Ann. Rev. Biochem.*, 61, 419–440.
41. Moore, C.L., Skolnik-David, H. and Sharp, P.A. (1986) *EMBO J.*, 5, 1929–1938.
42. Sheets, M.D., Stephenson, P. and Wickens, M.P. (1987) *Mol. Cell. Biol.*, 7, 1518–1529.
43. Sheets, M.D., Ogg, S.C. and Wickens, M.P. (1990) *Nucleic Acids Res.*, 18, 5799–5805.
44. Conzelmann, K.K. and Helftenbein, E. (1987) *J. Mol. Biol.*, 198, 643–653.
45. Maessen, G.D.F., Amons, R., Maassen, J.A. and Moller, W. (1986) *FEBS Lett.*, 208, 77–83.
46. Nomoto, M., Imai, N., Saiga, H., Matsui, T. and Mita, T. (1987) *Nucleic Acids Res.*, 15, 5681–5697.
47. Lai, Z.-C. and Childs, G. (1988) *Mol. Cell. Biol.*, 8, 1842–1844.
48. Gaertig, J. and Gorovsky, M.A. (1992) *Proc. Natl. Acad. Sci. USA*, 89, 9196–9200.