Nucleic Acids Research, Vol. 19, No. 17 4717–4724

Expression of the gene encoded by a family of macronuclear chromosomes generated by alternative DNA processing in *Oxytricha fallax*

Kevin R.Williams and Glenn Herrick*
Cellular, Viral and Molecular Biology, University of Utah School of Medicine, Salt Lake City, UT 84132, USA

Received April 26, 1991; Revised and Accepted August 5, 1991 GenBank accession nos M63171 – M63174 (incl.)

ABSTRACT

Hypotrichous ciliated protozoa, such as *Oxytricha fallax*, produce tiny chromosomes during generation of the transcriptionally active macronucleus. The 81-MAC family of macronuclear chromosomes is produced by alternative DNA processing, such that the chromosomes share a common region of 1.6 kbp. Transcription of a 1.3 kb mRNA from the common region has been analyzed. Transcription starts very near the telomere (34 bp), in a 23 bp region of pure A+T DNA. Polyadenylation sites are very near the other telomere (26 bp), also in a region of nearly pure A+T DNA. Three introns are clustered in the first third of the gene. Intron removal can follow polyadenylation, and the order of removal is not fixed. All three known sequence versions of the 81-MAC chromosomes are represented in the mRNA pool, with no evidence of any further versions. The A+T sequences surrounding the transcription starts and polyadenylation sites are conserved among versions. Introns have conserved 5′ and 3′ ends and a putative branch-point sequence (YYRAT), but otherwise are highly diverged and are AT-rich. A single long open reading frame, interrupted by the three introns, encodes a homolog of known mitochondrial solute carriers, and contains the codon TAA, which does not encode 'stop,' but a conserved glutamine; TAG appears also to encode glutamine. The results significantly enlarge the small data set of transcription start and polyadenylation sites, of intron features, and of translation signals for hypotrichs.

INTRODUCTION

Hypotrichous ciliated protozoa generate extremely small chromosomes during the development of the somatic or macronucleus, or MAC (1, 2). In *Oxytricha fallax* the number average size of MAC chromosomes is 2.50 kbp (J. Garrett and GH, unpublished). A variety of individual MAC chromosomes has been analyzed to date; in general a chromosome consists of telomeres at each end, a single transcription unit between, and very little else (3). The 81-MAC chromosome family is a set of three sizes of MAC chromosomes generated by alternative processing of copies of the germline DNA (4, 5, 6). The three chromosomes share a common region of 1.6 kbp. The smallest of the three (MAC III) consists solely of the common region, plus telomeres, and was found to carry at least a major part of a protein-coding gene. The MAC I chromosome consists of the common region plus an 'upstream' 3.0 kbp arm, and the MAC II chromosome consists of the common region plus a 1.4 kbp arm at the other end. Each of three distinct germline loci, or sequence versions, is independently processed to give rise to MAC I, II and III chromosomes; thus, the 81-MAC family consists of nine chromosomes, three version subfamilies of three chromosomes each (4, 5, 6, 10). A single 1.3 kb mRNA was detected by northern blotting of polyadenylated RNA from growing and from starved cells (5), and a large open reading frame was noted, which could encode a protein homologous to mitochondrial solute carriers. The bulk of the common region appeared to be devoted to this gene. However, the possibility of introns in turn left open the possibility that the transcription unit might lie in part outside the common region, on the MAC I or MAC II arm, although hybridization of northern blots with arm probes failed to detect the 1.3 kb transcript or any other. Only one hypotrich intron has been reported (7). Here we show that the 1.3 kb mRNA is derived from a 1.5 kbp transcription unit, which lies fully within the 1.6 kbp common region, and covers nearly all of it, and hence nearly all of the MAC III chromosome. The primary transcript has three introns interrupting the coding region for a homolog of mitochondrial solute carriers.

MATERIALS AND METHODS

DNAs

MAC chromosomes cloned in pMA plasmids are described by Cartinhour and Herrick (4) and Herrick et al. (5, 6). Oligomers, including dT_{30}, were synthesized with an Applied Biosystems instrument. See Fig. 1 for descriptions of oligomers specific to the 81-MAC common region; all have sequences which exactly match all three versions of the 81-MAC family sequence.

* To whom correspondence should be addressed

mRNA, and synthesis of cDNA isolation

Polyadenylated RNA was harvested from growing cells as previously described (5). First strand cDNA was synthesized by annealing 40 pmol of dT_{30} with 1.6 μg poly(A) selected RNA in 42mM KCl, 10mM tris-HCl, pH8.1, 1mM Na_2EDTA. The mixture was heated for 2 min at 95°, cooled slowly to room temperature, and adjusted to 0.2mM dNTPs, 30mM KCl, 10mM magnesium acetate, 50mM tris-HCl, pH8.1, 10mM dithiothreitol (DTT). Reactions were carried out at 42° for 30 min with 7.5 units of AMV reverse transcriptase (Life Sciences).

Polymerase chain reactions (PCR) and cloning of cDNA

Thermal cycler parameters used for all PCR sessions were: 1 min 95°, 1 min 55°, 3 min 72°, for 30 cycles with a final 7 min at 72°. Five different internal primer intervals, E to N, B to M, H to N, B to K, and B to G (see Fig. 1), were PCR amplified from first strand cDNA and used to obtain sequence of the mature mRNA. Oligomer F, which has its 3′ end (6 nt) within intron 3, and B, were used to amplify and sequence one additional interval from partially spliced mRNA still containing intron 3.

The strategy used for the amplification of 3′ ends is as described by others (8, 9). Session one of PCR paired primers dT_{30} and H on first strand cDNA. A 10% aliquot from the session one product was used as template for a second session of PCR. In session two, dT_{30} and I were used. Session two products were run on a 1.5% low melting point agarose (BRL) gel. A ~800 bp band was cut from the gel and one tenth of the gel slice was subjected to a third session of PCR, pairing dT_{30} with L. The session three PCR product was extracted with phenol-chloroform (1:1), precipitated from ammonium acetate and ethanol, and redissolved. DNA ends were made blunt by treatment with 5 units of T4 DNA polymerase (Promega) for 5 min at 37° in buffer recommended by Promega, then heated 75° 10 min. The DNA was cut with ClaI and ligated into the SmaI-ClaI cut pBluescript KSII+ vector (Stratagene) and used to transform DH5α cells (BRL). Clones containing the 3′ end of the cDNA were identified by the presence of a ~130 bp ClaI-EcoRI restriction fragment, and subsequent sequencing from a universal primer.

Primer extension

Oligomers were ^{32}P-5′ labeled with polynucleotide kinase and 7000 Ci/mM [γ-^{32}P] ATP (ICN). Sixty fmols of labeled oligomer were annealed with 1.6 μg of poly(A) RNA in 50mM tris-HCl, pH 8.5, 60mM NaCl, 10mM DTT, by heating 2 min at 95° then cooling slowly to room temperature. Extension reactions were performed in 6mM magnesium acetate, 0.18mM each dNTP, 25mM tris-HCl, pH 8.5, 30mM NaCl, 5mM DTT, with 10 units of AMV reverse transcriptase for 30 minutes at 45°. In the dideoxy-termination reactions all conditions remained the same except five fold more poly(A) RNA (8 μg) was used, and the d:ddNTP mix was 0.28mM each dNTPs and 0.57mM of either ddATP or ddTTP.

Sequencing

PCR products representing the interval between primers B and N were isolated from 2% agarose gels and purified away from the agarose using Geneclean (Bio 101). Various of the oligomers B to N were used as sequencing primers, so that the entire region was sequenced; the sequence was verified by comparison to macronuclear DNA sequences. Cloned 3′ end DNAs were isolated from small (6 ml) cultures of transformed DH5α cells. Dideoxy-termination sequencing was performed using the Sequenase kit, version 2.0 (United States Biochemical) and [α-^{35}S] dATP (>1000 Ci/mM, Amersham).

In vitro transcription

The vB macronuclear chromosome III cloned in pMA83 (4) was moved into the pBluescript KS+ vector, oriented so that T7 RNA polymerase transcribes in the same sense as the chromosome is transcribed *in vivo*. The plasmid was linearized by cutting at the HincII site at far end of the insert and used as the transcription template. RNA (5−10 μg) was synthesized using the 'transcription protocol #2' (Promega).

GenBank accession numbers. Genomic sequences previously submitted: M13029−042, M25391; new sequences: M63171−74.

RESULTS AND DISCUSSION

The 1.6 kbp region common to the three chromosomes of the 81-MAC family is known to encode at least part of a 1.3 kb mRNA (5). DNA complementary (cDNA) to that mRNA now has been synthesized and sequenced, the mRNA 5′ end and polyadenylation sites have been determined, and three introns have been identified. **Figure 1** presents a map of the region and summarizes our findings.

Transcription of all three versions of 81-MAC sequences

Three independent sequence versions of the 81-MAC family chromosomes have been identified in the *O. fallax* genome (4, 5, 6, 10). Each version, vA, vB and vC, is represented as a set of three MAC chromosomes. Genomic DNA analyses showed no evidence for any further versions. Sequencing of the common regions of the three versions showed that they differ primarily by scattered base substitutions, and that areas of high conservation are interspersed with areas of larger divergence; overall, the versions differ from each other by from 3.2% to 4.9%. To determine which versions are represented in mRNA, cDNA has been sequenced in bulk to look for version-diagnostic bands. Pairs of exon oligomers were designed as polymerase chain reaction (PCR) primers, and were used to amplify first strand cDNA. Products of these reactions then were sequenced. Nucleotides which are diagnostic for a given version were used to determine which versions are present. An example is shown in **Figure 2**; note one band diagnostic for vA, two for vB, and one for vC. These results show that all three versions of macronuclear DNA are transcribed. The entire cDNA has been sequenced, either in bulk, as shown here, or as clones of 3′ ends (see below). All expected version-diagnostic bands were seen (16 vA, 12 vB and 16 vC; not shown). Other bands that would indicate the transcription of an unknown fourth version were not seen.

Transcription initiation sites

The position of the 5′ end of the mRNA was determined by primer extension, using three different primers (A, C, D; Fig.1A). The results (not shown) consistently placed the 5′ end of the mRNA near the boundary of the common region, taking into account introns that intervene (see below) between the mRNA 5′ end and the binding sites for primers C and D. It remained possible that an intron exists between primer A and the 5′ end of the mRNA. This possibility was tested by sequencing the mRNA in that region. Sequence was obtained by extending primer A with reverse transcriptase in the presence of dideoxy

nucleotides (**Fig. 3**). Comparison of genomic sequence to that of the mRNA showed there were no additional introns. Two reverse transcriptase stops were seen (Fig. 3B), a strong stop and a weaker stop, 16 and 18 nt from the primer A binding site, and separated by a single nucleotide that is different between versions (nt 59; Fig. 1B). It is conceivable that secondary structure within the RNA might cause the polymerase to stall, resulting in the stops. However, the region is extremely A+T rich and unlikely to form unusually strong secondary structures. Also, polymerase proceeded through this region on a control RNA, produced by *in vitro* transcription of a cloned segment carrying the entire common region (Fig. 3A). The two stops seen by primer extension on the mRNA sample are taken to indicate the positions of two transcription start sites for the mRNA.

Tetrahymena transcription often starts immediately after a T, at an A (Csank and Martindale, personal communication). In the present case, the upstream site for all three versions is TA (nts 57 and 58, Fig. 1B), as are the downstream sites of vB and vC (nts 59 and 60, Fig. 1B). However, the downstream site of vA is AA (nts 59 and 60, Fig. 1B); it is possible that vA transcripts only start at the upstream site. The region surrounding the start sites is fully conserved between versions, except for the single vA difference between the two sites. A 23 bp section containing the start sites is purely A+T. Initial attempts to map the mRNA ends by nuclease protection failed; in retrospect, the fact that 5′ ends (and 3′ ends, see below) lie in regions of pure A+T likely explains those failures. On the macronuclear chromosome cloned in pMAμ1A1 (6), the first transcriptional start site is only 34 bp from the telomere sequence, which leaves little space for upstream signals. The sequence TATAAA was identified, which conforms to the consensus of the eukaryotic TATA element (Fig. 1B). However, it is found only 6 bp upstream of the transcriptional start, casting doubt on its significance. Furthermore, the match might well be accidental, since the 5′ non-transcribed region is rich in A+T (91% for the 34 bp of pMAμ1A1), and the sequence of TATA elements can vary widely (11). No potential CCAAT elements are seen, unlike in some other hypotrich and *Tetrahymena* genes (12).

Polyadenylation sites

In order to locate polyadenylation sites, clones were generated from PCR-amplified cDNA (Materials and Methods). Primary cDNA amplification was performed using the primer dT_{30} paired with an exon primer (8). Improved specificity was obtained by secondary and tertiary amplification sessions (9), using primers progressively closer to the 3′ end. Plasmid clones of 3′ end cDNA

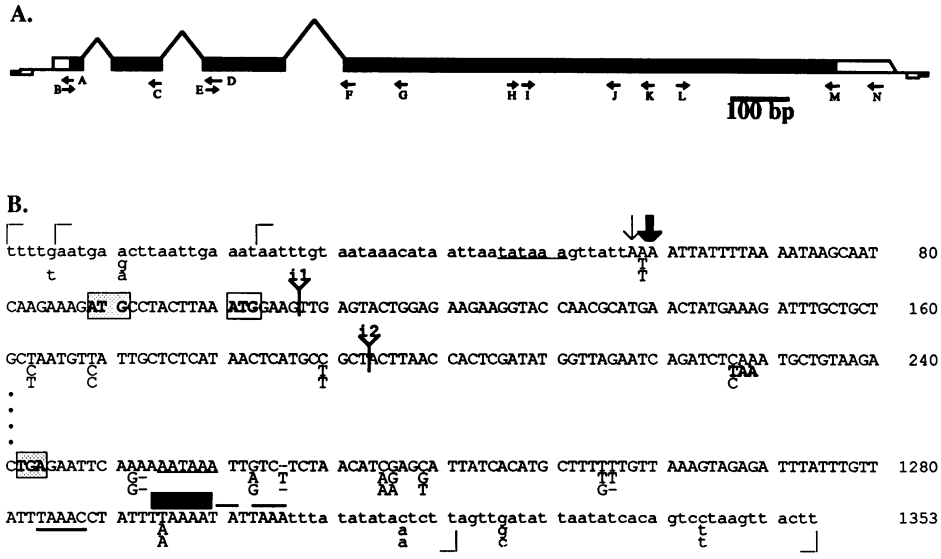

Figure 1. Transcription unit of the region common to 81-MAC chromosomes. **A**. Map. The genomic region represented extends from the telomere (terminal box) of the chromosome carried in pMAμ1A1 to the telomere of pMA81. Previously published maps have the opposite polarity. Transcription is from left to right. Three introns connect four exons (boxes). The long open reading frame is represented by filled-in exon regions. Oligomers are represented as 3′-tipped arrows, and labeled with single-letter names. Following is a listing of oligomers, giving the full name, in parentheses, followed by the sequence, and then the coordinates on the sequence shown in Fig. 1B: A, (VHO) dGGCATCTTTCTTGATTGC (93−76); B, (VHO′) dGCAATCAAGAAAGATGCCTAC (76−96); C, (PE-C) dGCATGAGTTATG-AGAGCAAT (189−170); D, (PE-B) dAGATCTGATTCTAACCATATCGAGTGG (226−200); E, (PE-B′) dCACTCGATATGGTTAGAATCAG (201−222); F, (L1) dGAAGAATGTGTCGAAACTATAC (361−345 + 6nt of intron 3); G, (83s # 3) dGGACGAGCAACTCTTCTGGG (453−434); H, (oRG) dAGAGGTGCT-GGTGCCAAC (632−649); I, (83s # 2) dGCTTGCAGCTATTTGTTCATC (658−678); J, (oDM′) TAGTCTTGTTCTGATCATATC (823−803); K, (oGM′) dT-TGAAGCAATCAATCATGCC (885−866); L, (T1-target′) dACTTTGGATCATTCTATGCAG (927−947); M, (LCR2) dCTCAGATAATTGCGAGATTTTC (1205−1184); N, (LCR1) dATAACAAATAAATCTCTACTTTAAC (1279−1255). **B**. Sequences at the beginning and end of common region gene: sequences of cDNAs and common region genomic flanks. Sequences from position 241−1200 have been omitted; full sequences are available in GenBank via accession numbers M63171−74. The sequence shown is that of the version A (vA) macronuclear chromosome III carried in pMA83s, without telomeric repeats, and without introns (see Fig.4 for intron sequences). Letters below each line of sequence represent vB and vC nucleotides, where they differ from vA. In addition, the vB codon TAA (227−229) is fully listed, and encodes glutamine (see text). cDNA sequences are in upper case, and non-transcribed subtelomeric flanks are in lower case. Positions of introns 1 and 2 are indicated by 'Y' symbols; the position of intron 3 is between nucleotides 345 and 346 (not shown). Major and minor transcription starts determined by primer extension are indicated by thin and thick vertical arrows above As at positions 58 and 60 (see text and Fig. 3). Three regions of polyadenylation, one major and two minor, are indicated by thick and thin bars over positions 1294−1306. Potential transcription initiation and polyadenylation signals are underlined. The large open reading frame start and stop codons are boxed and shaded, and an in-frame ATG near the beginning is boxed. Positions of 5 chromosome-breakage/telomere-additions represented by various cloned macronuclear chromosomes are indicated before positions 1, 6 and 24 (pMAs 83s, 83 and μ1A1, resp.) and after positions 1321 and 1353 (pMAs 81 and 83s).

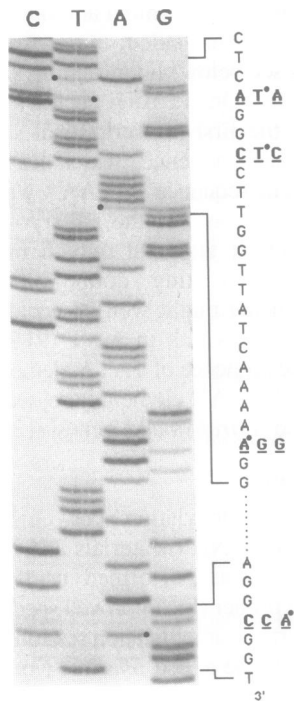

Figure 2. Representation of all three versions in mRNA. cDNA between oligomers E and K (see Fig. 1) was PCR amplified and sequenced. Shown is an autoradiogram of the sequencing gel of reactions primed with G, which is complementary to the mRNA sequence. So that the figure indicates the mRNA sequence, the gel lanes have been labeled with the complements of the ddNTP actually used in the sequencing reactions. Portions of the sequence are listed vertically. Underlined letters mark positions where the sequence is different between versions vA, vB, and vC respectively. These version differences correspond to nucleotide 256, 259, 274 and 319 in Fig. 1B. Nucleotides diagnostic for a single version are marked with black dots.

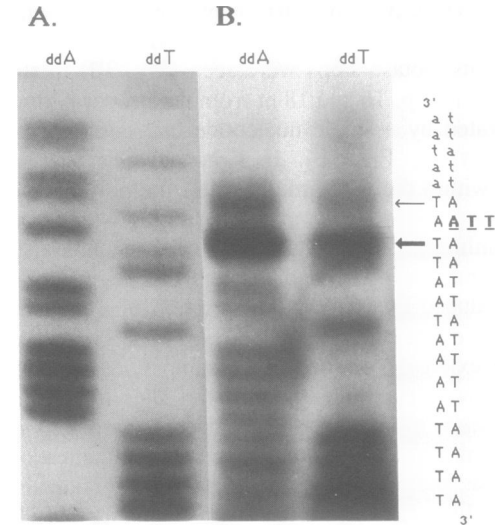

Figure 3. Primer extension sequencing of mRNA 5′ ends. End-labeled primer A (see Fig. 1) was extended with reverse transcriptase in the presence of ddATP or ddTTP. Autoradiograms of sequencing gels are shown. **A.** The template was transcribed *in vitro* by T7 RNA polymerase across the entire cloned vB chromosome III of pMA83. **B.** Poly(A) RNA template. Both strands of the corresponding sequence are listed vertically at the right, and correspond to nt 53−72 of Fig. 1B. Upper case letters represent nucleotides of mRNA, and lower case letters represent nucleotides present only in the *in vitro* transcript. Two stops are indicated by arrows; the thin arrow indicates the weaker band, and the bold arrow indicates the stronger band. Note that the sequence has no Gs or Cs: only the two largest bands were seen in another experiment (not shown) where extensions were performed in the presence of dATP, dTTP, dCTP, and ddGTP, but no dGTP, or in the presence of dATP, dTTP, dGTP, and ddCTP, but no dCTP. These strong stops correspond to the similarly marked transcription starts in Fig. 1B. The position between the stops shows no band, presumably because it varies from version to version. The nucleotide at that position in each version is indicated by an underlined letter in the right-hand strand, in the order vA, vB, and vC.

were derived from the tertiary PCR product. The cDNA inserts were sequenced to establish version identity, and to locate sites of polyadenylation (Materials and Methods).

All three versions are represented in the inserts, again demonstrating that each version is transcribed. Of 49 clones examined, most were vB and vC (29 and 10, respectively). Only one was vA. The remaining 9 clones contained mixed version inserts (3 vBC, 5 vAB and 1 vAC). We believe these were derived from heteroduplexes formed in late rounds of PCR. In other work we have demonstrated the formation of such heteroduplexes between strands of different versions and between micronuclear and macronuclear sequences (K.W. and G.H., unpublished).

The clone inserts ended with poly(A) runs that varied in length from 15 to 87 bp. These values are unlikely to represent mRNA tail length, because during PCR the dT_{30} oligomer can anneal in many locations along the poly(A) tail, creating a mixture of different lengths. The encoded length of the mRNAs is 1235 to 1248 bp. The apparent mRNA length deduced from northern blots (5) is 80−100 nt larger, suggesting the poly(A) tails are about that size.

Polyadenylation usually occurs (44 of 49 clones) after one of 5 contiguous As (Fig. 1B). It is unknown which As are templated and which are added post-transcriptionally. Similarly, the minor sites (5 clones) cannot be placed unambiguously, but are just downstream of the major cluster of sites (Fig. 1B). The sites are in a 27 bp region that is purely A+T and fully conserved between versions, except for one difference in vA (Fig. 1B). It is possible that during PCR amplification, mis-annealing of the dT_{30} oligomer to short genomic A runs could generate primer-templated poly(A) runs 5′ of the actual site for polyadenylation. Consistent with this possibility, conventional signals for polyadenylation (13) are not found at the appropriate locations, although a potential cleavage signal AATAAA is seen ~80 bp upstream of the polyadenylation sites (Fig. 1B). However, polyadenylation sites have been identified for four other hypotrich genes, and these were determined independently of PCR (14, 15, 16). These three genes also lack conventional polyadenylation signals. They do each contain related pentanucleotides TAAAC, TAAAC, AGAAC and TGAAC, respectively, at, or just 5′, of their polyadenylation site (consensus: tRAAC). In the present case, TAAAC is also seen just 5′ of the putative first site of polyadenylation (Fig. 1B). The recurring coincidence of the tRAAC sequence with the polyadenylation site, and lack of other conventional signals, bolster the suggestion that it may play a signaling role in hypotrich 3′ end RNA processing (14, 15).

The telomere of the chromosome cloned in pMA81 is at most 26 bp from any polyadenylation site (see Conclusion). The shortest ciliate 3′ non-transcribed subtelomeric region previously reported is 195 bp (15). Non-translated 3′ regions of 37, 53 and 78 bp have been reported for two other hypotrich chromosomes (compared to 115 bp for the chromosome of pMA81; see below),

```
i1
vA  agGTAAATTTATCAATAAAATATCGAATCA                                    TCGAT  TTATAAAATTTATTTTTAG
vB  agGTAAATTTATCAATAAAATATCGAATCA                                    TCGAT  TTATAAAATTGCTTTTTAG
vC  agGTAAATTTATCAATAAAATATCGAATCA                                    TTGAT  TTCTAAAATTTCGTTTTAG
      |||| || ||                                                     |*|||    *       *** | |||
i2    |||| || ||                                                     |||||                 | |||
vA  ctGTAAGTTAATACCCTAAAGCACA-ATATACTATGCTTTGTGATAT                   TTAATCATATTCTTAATGTAATATAG
vB  ctGTAAGTTAATATCCTTTAGCACAGTTTCTATCTTATTAGCAATGA                   TTAATCATATTCCTAATGTAATATAG
vC  ctGTAAGTTAGTATTCTATAGCACATTTTCTATCTTATTAGCAATGA                   TTAATCATATTCTAAATGTAATATAG
      |||| || *| **  **      ** **** * **  * **  **                  |||||       **        | |||
i3    |||| || ||                                                     |||||                 | |||
vA  gcGTAAGTGTATAAACAAAATTTGGGTTATCCAAAAACATCTCTCAATTTTATAATTGAGGAATTGCAATAAATTTAGCTAAT       TTATATTGGTATAG
vB  gcGTAAGTTTATAAACAAAATTTGGGTTATCCAAAAAGCTCTCTCAACCTTATAATTGAGGAATTGCAATAAATTTAGCTAAT       TTATATTGGTATAG
vC  gcGTAAGTGTATAAACAAAATTTGGCTTATATAAAGTTCTCTCTCAACATTATAATTGAGGAATTGCAATAAATTTAGCTAAT       TTATATTGGTATAG
          *                 *    **   ****        **
con   GTAA Tt aT                                                     YtRAT               T TAG
```

Figure 4. Conserved intron sequence features. Intron sequences are shown in upper case letters. Two nucleotides of the 5′ flanking exon are shown in lower case letters. Vertical lines connect shared nucleotides at intron ends and at a possible branch point consensus. Consensus nucleotides listed below are in upper case if the match is perfect and in lower case if the match is better than 6 of 9. Underlined sequences show potential TAG acceptor sites which are not used, indicating the branch points must be downstream. In the consensus sequence of the putative branch points, Y means C or T, and R means A or G. Locations where nucleotides differ between versions are marked with asterisks. Sequence differences between versions: i1, vA:vB 3.8%, vA:vC 7.7%, vB:vC 7.7%; i2, vA:vB 25%, vA:vC 27%, vB:vC 8.5%; i3, vA:vB 5.3%, vA:vC 9.5%, vB:vC 8.4%. Intron A+T compositions: i1, vA 85%, vB 81%, vC 81%; i2, vA 76%, vB 75%, vC 78%; i3, vA 76%, vB 74%, vC 76%.

and these chromosomes might have 3′ non-transcribed regions even shorter than 26 bp (17, 18).

Three introns

An internal section of the cDNA was amplified by PCR with the primers B and G. Several discrete sized products, all reverse transcriptase dependent, were seen by agarose gel electrophoresis (not shown). The major product was ~225 bp smaller than the corresponding product templated by macronuclear genomic DNA. It was gel purified and sequenced. Comparison its sequence to that of genomic DNA revealed three introns (**Fig. 4**). The exons and introns become progressively larger as one scans along the gene (see Fig. 1A). Intron 1 (i1) is 52 bp, i2 is 71 bp (70 nt for vA), and i3 is 95 bp. Exon 1 (e1) is 48 to 50 bp, e2 is 86 bp, e3 is 152 bp, and e4 is 949−961 bp.

Note that the introns are clustered at the beginning of the gene, reminiscent of many genes in *Saccharomyces cerevisiae*. Fink (19) suggested this might have resulted from recombination between genes and cDNAs lacking 5′ sequences. Thus, cDNAs generated from macronuclear transcripts might have recombined with the corresponding germline micronuclear gene.

Each intron is poorly conserved from version to version, compared to the 2% dissimilarity between the versions of the protein coding region, which they interrupt (see below). The average dissimilarity between versions is 6% for the i1's, 20% for the i2's, and 7% for the i3's. The introns have high A+T compositions, greater than 74% (Fig.4 legend). They contain the eukaryotic invariant 5′ GT and AG 3′ splice site nucleotides. A consensus was extended to include other conserved nucleotides internal to the invariant 5′ GT and AG 3′ (Fig.4). The extended consensus sequences and high A+T composition are similar to those found in the introns of other eukaryotes, including *Tetrahymena* (20), and to that of the only other known hypotrich intron (7). A possible branch point sequence YTRAT, consistent with the mammalian branch site consensus YTRAY (21), is seen in all introns except vA and vB of i1, which have the consensus YCRAT (Fig. 4).

PCR amplification of cDNA from e1 to e4 (primers B to G) not only produced the major intron-less product described above, but also produced a heterogeneous set of minor, larger products. These were derived from intron-containing but polyadenylated RNAs. The heterogeneous cDNA products were fractionated into

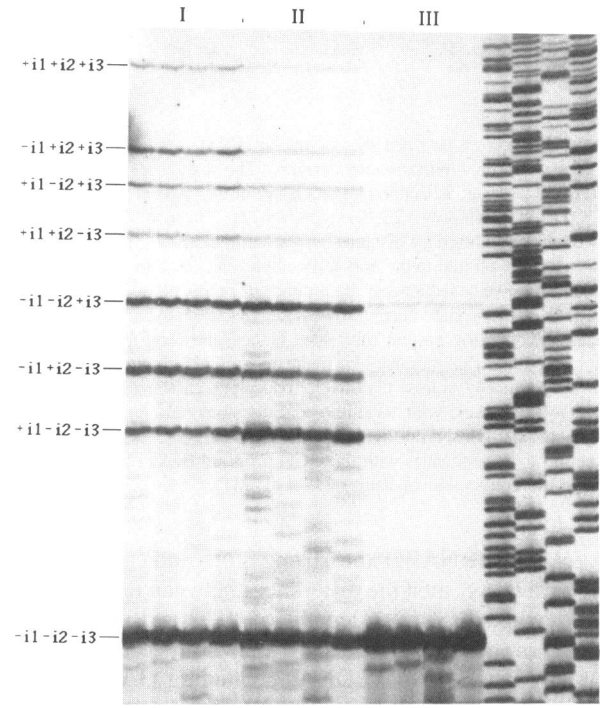

Figure 5. cDNAs from incompletely spliced, polyadenylated mRNAs. cDNA PCR amplified between primers B and G (Fig. 1A) was purified into size fractions I, II, and III (large to small) and subjected to sequencing reactions primed with oligomer G. Strong stops are marked and taken to represent ends of cDNA templates. The interpretation of the structure of each cDNA is indicated at the left, showing the presence (+) or absence (−) of i1, i2 or i3. The four righthand lanes serve as size standards. The length of the interval between each adjacent pair of bands was compared to the predicted interval. The seven intervals, from top to bottom, are listed as observed, then predicted and the discrepancy: 52 vs. 50 bp (Δ = 2 bp), 19 vs. 18 bp (Δ = 1 bp), 24 vs. 23 bp (Δ = 1 bp), 28 vs. 26 bp (Δ = 2 bp), 24 vs. 24 bp (Δ = 0 bp), 19 vs. 18 bp (Δ = 1 bp), 52 vs. 51 bp (Δ = 1 bp).

three size classes, which were used as templates for sequencing reactions with primer G. Eight strong stops were observed (**Fig. 5**). If the three introns are removed independently of each other, then cDNAs of eight different lengths would be expected, seven derived from incompletely spliced mRNAs, and one from

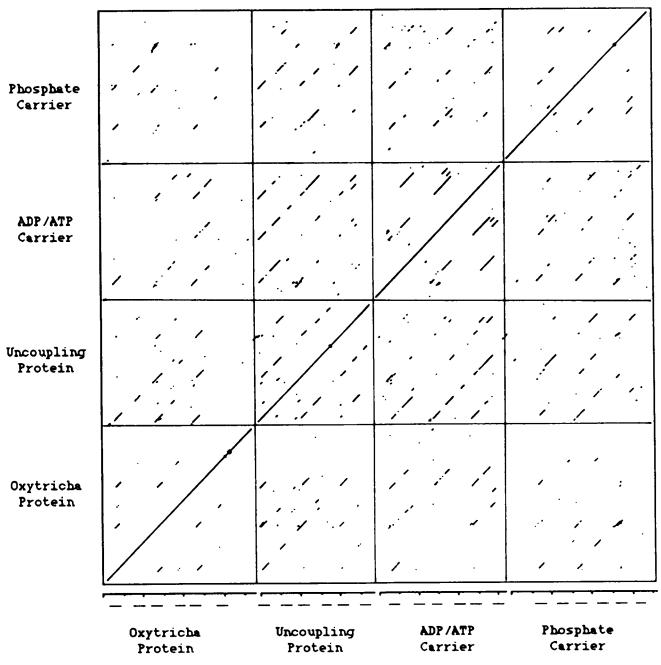

Figure 6. Homology between the *Oxytricha* HMSC amino acid sequence and sequences of three bovine solute carriers. The sequences (31,36, 37) were compared by the diagonalization plot procedure of the GCG programs COMPARE and DOTPLOT (window=20, stringency=12). The comparison matrix used scores relative evolutionary distance between amino acids (38, 39). The number of points ('dots') plotted in the non-self comparions gives an indication of extent of similarity between sequences; the number of points plotted in all the non-self comparions is: phosphate carrier row, 117, 184, 201; ADP/ATP carrier row, 147, 234; Uncoupling protein row, 159. Hence, the *Oxytricha* HMSC protein is apparently somewhat more similar to the uncoupling protein and the ADP/ATP carrier (159 and 147 points) than to the phosphate carrier (117 points). Below the plots are size axes, with 50 amino acid divisions, and lines representing the positions of six conserved hydrophobic regions I-VI, determined by local alignments of the sequences with the GCG program BESTFIT (not shown) and following the convention of Runswick et al. (31).

fully spliced mRNA. If the strong stops indicate the ends of template cDNAs, then the differences between template lengths can be predicted from the lengths of i1, i2 and i3. Sizes of the predicted products correspond, within 2 nt, to those observed (Fig.5). A subset of four cDNAs containing i3 were amplified using primers F and B; the smallest product was gel-purified and sequenced. It contains i3 but precisely lacks i1 and i2, confirming the conclusion. The three introns apparently are not spliced in a fixed order and polyadenylation can precede splicing, as is known in other systems (for a recent discussion see 22), but not previously in ciliates.

Protein coding region and translation of TAA and TAG codons

The cDNA sequence has an apparent long, 371 codon open reading frame (ORF), terminated by TGA. It barely begins in e1 (6 codons), with the bulk in the final three exons (Fig. 1). The ORF has several codons conventionally read as 'stop:' the vA ORF has 5 TAAs and 1 TAG, vB has 6 TAAs and 1 TAG, and vC has 5 TAAs. In a variety of other ciliates TAA and TAG do not encode 'stop' but probably encode glutamine, and TGA encodes 'stop' (review, 23). However, there are exceptions (see below). In *Oxytricha* species, including *O. fallax*, TGA encodes 'stop' for several genes studied, but no information exists about TAA and TAG translation. Various features of the present ORF allow us to draw conclusions about the TAA and TAG translation in *O. fallax*.

Four points argue that the entire ATG-to-TGA region encodes amino acids and therefore that TAA and TAG do not encode 'stop.' (i.) RnY codons are used preferentially in that frame, which is true for the protein coding frames of many other genes (24): the entire putative ORF has 35% RnY codons, compared to 20 and 21% in the other two frames; codon usage is biased all the way to the TGA (5). (ii.) Similarly, the entire open frame has a relatively high G+C content (40%) compared to flanking sequences and introns (<26%), and this bias also extends to the TGA (5). (iii.) The full length of the ORF is highly conserved between versions, only 2% diverged at the nucleotide level. At the amino acid level vA and vB are identical, and vC differs by two conservative changes (K for R and A for V, not shown; 5); this is in contrast to the sequences outside the ORF (see above and below). (iv.) The putative protein encoded is homologous to known proteins, the mitochondrial solute carriers, along nearly its full length (see below), including residues encoded by TAA codons.

TAA translation differs in different ciliates. In the holotrichs *Tetrahymena* and *Paramecium* it encodes glutamine, as it does in the oxytrich *Stylonychia lemnae* (23). However, in various *Euplotes* species it encodes 'stop' (16, 18). Two facts from the present work argue that TAA encodes glutamine in *O. fallax*. (i.) Codon 47 is TAA in version B (nts 227–229, Fig. 1B); the corresponding amino acid is glutamine in all members of the mitochondrial solute carrier superfamily (see below). (ii.) Codon 47 is a conventional glutamine codon in vA and vC; as noted, the proteins encoded by the three versions are otherwise nearly perfectly conserved.

TAG translation has not been reported in any hypotrich, although it encodes glutamine in the holotrichs (23). Codon 322 of the vC ORF is CAG, a conventional glutamine codon; the corresponding codon in vA and vB is TAG (not shown), arguing that TAG encodes glutamine in *O. fallax*.

In summary, TAA encodes glutamine in two oxytrichid hypotrichs, *O. fallax* and *S. lemnae*, and in holotrichs, but encodes 'stop' in euplotid hypotrichs, re-enforcing the contradiction with the taxonomic splitting of hypotrichs from holotrichs (16, 18). TAG appears to encode glutamine in *O. fallax*, as it does in holotrichs; it remains to be seen if it also encodes glutamine in other oxytrichid hypotrichs, and what it encodes in euplotid hypotrichs.

Translation start and stop

The 5′ transcribed but not translated region is 31 bp long. It has a high A+T composition (84%). It is conserved among versions, except for the vA difference that separates the transcription starts (Fig. 1B). Translation presumably begins at the first AUG on the mRNA (25), although there is a second in-frame AUG just four codons downstream of the first AUG (Fig. 1B). Contexts of the two AUGs—AAG AUG C and UAA AUG G—do not conform to the consensus translation start sequence, AAA AUG G, which has emerged for a biased set of ciliate genes (12, 26). At the 3′ end, a short non-conserved region (43 or 45 bp) follows the UGA stop codon and precedes the conserved 3′ region of polyadenylation (above). It is poorly conserved in both length and sequence (13%-15% dissimilar between versions). The entire 3′ transcribed but not translated region (90–101 bp) is rich in A+T (75%-80%).

Homolog of mitochondrial solute carriers

Database searches revealed a similarity between the encoded protein and members of the superfamily of nuclear-encoded solute carrier proteins of the inner mitochondrial membrane (27), the best characterized of which are the family of the ADP/ATP carriers (28). The superfamily also includes a phosphate carrier and an uncoupling protein, the so-called mitochondrial solute carrier analogs (29), a novel member from *Caenorhabditis elegans* (30), and likely several more (see 31). Figure 6 shows diagonalization plots establishing homology between three members of known function, all bovine, and the *Oxytricha* common region-encoded protein, hereafter referred to as the homolog of mitochondrial solute carriers or HMSC. In addition to modest central diagonals, these plots show secondary diagonals reflecting an ancient triplication of a ~100 residue repeat, a character the HMSC shares with the carriers. The regions of strongest similarity roughly coincide with hydrophobic membrane spanning regions (27, 28); each triplication is bracketed by a pair of these (Fig. 6 bottom). Hydrophobicity profiles (not shown) demonstrate that the HMSC shares this character with mitochondrial solute carriers. Local alignment of amino acid sequences demonstrates (not shown) that regions of high similarity carry residues which are highly conserved between all superfamily members. For instance, in the sequence **PLD**MVR**IRSQ**, near the end of the first hydrophobic region of the HMSC, and in corresponding sequence **PLD**TAKV**RLQ** in the bovine uncoupling protein, the terminal glutamine residue (**Q**) is found in all members of the superfamily (27, 28, 29, 30, 31). It is encoded by the TAA codon 47 of the HMSC (see above).

Figure 6 shows the HMSC is distinctly different from any of the known solute carriers. (See Fig. 6 legend for numerical measures of similarity; plots, not shown, of the HMSC against the mitochondrial solute carrier 'analogs' (29), and the *C. elegans* homolog (30), demonstrate low extents of similarity, comparable to those seen in Fig. 6.) What is the evolutionary relationship between the HMSC and the known carriers? Is it a new member of the superfamily? Or, is it simply a highly diverged ortholog of one of the already known genes? ADP/ATP carrier orthologs from plants, yeast and chordates are all highly similar to one another (28), whereas the HMSC is only modestly similar to any superfamily member in any of those taxa (not shown). Ciliates, plants, fungi and chordates are felt all to have diverged from one another more or less at the same time, judged from evolutionary distances between small subunit ribosomal RNA sequences (32, 33, 34). This implies that the HMSC gene must have diverged from the known carrier genes well before that radiation, and therefore is not orthologous to any of them. It is worth noting, however, that orthologous proteins which are highly conserved amongst the plants, fungi and chordates may nonetheless evolve especially rapidly in the ciliates, as evidenced by ciliate histones H3 and H4 (35), and actins and tubulins (M.Sogin, pers. comm.).

CONCLUSION

The MAC III chromosome of the 81-MAC family is almost entirely devoted to a transcription unit (Fig. 1), leaving much less than 100 bp of non-transcribed, non-telomeric sequences which could be dedicated solely to other functions, such as transcription and DNA replication elements. This was not apparent from the initial comparison of mRNA size and chromosome size, due to the existence of introns; as more hypotrich transcription units are fully mapped, the number of such instances should grow. MAC development mechanisms have evolved for editing away massive lengths of sequences unessential for gene expression. It is curious that within the transcription unit, the poorly conserved 3′ untranslated region and the introns nonetheless persist in the MAC.

We have found no ready explanation for the function of the larger chromosomes in the 81-MAC family, since we have detected no transcripts from their arms, either by northern analysis (5), nor now by mapping the 1.3 kb mRNA. Also, PCR has failed to detect transcripts crossing from the common region into or from either arm (unpublished). It is conceivable that the arms provide cis elements required for expression of the common region gene under special circumstances, or that the arms themselves carry alternative exons or separate genes expressed under special circumstances. Indeed, it cannot be ruled out that the 1.3 kb mRNA is derived solely from MAC I or MAC II, and that MAC III chromosomes are transcriptionally inactive, being so lean. A variety of other MAC chromosomes are, however, similarly lean, and do generate mRNAs (3).

ACKNOWLEDGEMENTS

Support for this work was provided by the NIH grant GM25203. We thank C.Csank and D.Martindale, and M.Sogin, for communication of unpublished results, and E.Ehrenfeld for critically reading of the manuscript.

REFERENCES

1. Klobutcher,L.A. and Prescott,D.M. (1987) In Gall,J.G. (ed.), The Molecular Biology of Ciliated Protozoa. Academic Press, Inc., Orlando, Florida, pp. 111–154.
2. Yao,M.-C. (1989) In Berg,D.E. and Howe,M.M. (eds.), Mobile DNA. American Society for Microbiology, Washington, D.C., pp. 715–734.
3. Harper,D.S. and Jahn,C.L. (1989a). *Gene*, **75**, 93–107.
4. Cartinhour,S.W. and Herrick,G.A. (1984) *Mol. Cell. Biol.*, **4**, 931–938.
5. Herrick,G., Hunter,D., Williams,K. and Kotter,K. (1987a) *Genes Dev.*, **1**, 1047–1058.
6. Herrick,G., Cartinhour,S.W., Williams,K.R. and Kotter,K.P. (1987b) *J. Protozool.*, **34**, 429–434.
7. Hicke,B.J., Celander,D.W., MacDonald,G.H., Price,C.M. and Cech,T.R. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1481–1485.
8. Frohman,M.A., Dush,M.K. and Martin,G.R. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 8998–9002.
9. OHara,O., Dorit,R.L. and Gilbert,W. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 5673–5677.
10. Hunter,D.J., Williams,K., Cartinhour,K. and Herrick,G. (1989) *Genes Dev.*, **3**, 2101–2112.
11. Singer,V.L., Wobbe,C.R. and Struhl,K. (1990) *Genes Dev.*, **4**, 636–645.
12. Brunk,C.F. and Sadler,L.A. (1990) *Nucleic Acids Res.*, **18**, 323–329.
13. Humphrey,T. and Proudfoot,N.J. (1988) *Trends in Genet.*, **4**, 243–245.
14. Conzelmann,K.K. and Helftenbein,E. (1987) *J. Mol. Biol.*, **198**, 643–653.
15. Helftenbein,E. and Müller,E. (1988) *Curr. Genet.*, **13**, 425–432.
16. Miceli,C., La Terza,A. and Melli,M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 3016–3020.
17. Ribas-Aparicio,R.M., Sparkowski,J.J., Proulx,A.E., Mitchell,J.D. and Klobutcher,L.A. (1987) *Genes Dev.*, **1**, 323–336.
18. Harper,D.S. and Jahn,C.L. (1989b) *Proc. Natl. Acad. Sci. USA*, **86**, 3252–3256.
19. Fink,G.R. (1987) *Cell*, **49**, 5–6.
20. Csank,C., Taylor,F.M. and Martindale,D.W. (1990) *Nucleic Acids Res.*, **18**, 5133–5141.
21. Gelfand,M.S. (1989) *Nucleic Acids Res.*, **17**, 6369–6382.
22. LeMaire,M.F. and Thummel,C.S. (1990) *Mol. Cell Biol.*, **10**, 6059–6063.
23. Martindale,D.W. (1989) *J. Protozool.*, **36**, 29–34.
24. Shepherd,J.C.W. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 1596–1600.
25. Kozak,M. (1978) *Cell*, **15**, 1109–1123.

26. Horowitz,S., Bowen,J.K., Bannon,G.A. and Gorovsky,M.A. (1987) *Nucleic Acids Res.*, **15**, 141–160.
27. Aquila,H., Link,T.A. and Klingenberg,M. (1987) *FEBS Lett.*, **212**, 1–9.
28. Klingenberg,M. (1989) *Arch. Biochem. Biophys.*, **270**, 1–14.
29. Zarrilli,R., Oates,E.L., McBride,O.W., Lerman,M.I., Chan,J.Y., Santisteban,P., Ursini,M.V., Notkins,A.L. and Kohn,L.D. (1989) *Mol. Endochrinol.*, **3**, 1498–1508.
30. Wang,H. (1989) Ph.D. thesis, University of Utah.
31. Runswick,M.J., Powell,S.J., Nyren,P. and Walker,J.E. (1987) *EMBO J.*, **6**, 1367–1673.
32. Sogin,M.L., Gunderson,J.H., Elwood,H.J., Alonso,R.A. and Peattie,D.A. (1989) *Science*, **243**, 75–77.
33. Patterson,C. (1990) *Nature*, **344**, 199–200.
34. Douglas,S.E., Murphy,C.A., Spencer,D.F. and Gray,M.W. (1991) *Nature*, **350**, 148–151.
35. Brunk,C.F., Kahn,R.W. and Sadler,L.A. (1990) *J. Mol. Evol.*, **30**, 290–297.
36. Costeilla,L., Bouillaud,F., Forest,C. and Ricquier,D. (1989) *Nucleic Acids Res.*, **17**, 2131.
37. Powell,S.J., Medd,S.-M., Runswick,M.J. and Walker,J.E. (1989) *Biochemistry*, **28**, 866–873.
38. Dayhoff,M.O., Barker,W.C. and Hunt,L.T. (1983) *Meth. Enzymol.*, **91**, 524–545.
39. Gribskov,B. and Burgess,R. (1986) *Nucleic Acids Res.*, **14**, 6745–6763.