MOLECULAR AND CELLULAR BIOLOGY, Nov. 1990, p. 6091–6096 Vol. 10, No. 11
0270-7306/90/116091-06$02.00/0

Molecular Characterization of *SerH3*, a *Tetrahymena thermophila* Gene Encoding a Temperature-Regulated Surface Antigen

M. MEHRDAD TONDRAVI,[1]†* REBECCA L. WILLIS,[2]‡ HAROLD D. LOVE, JR.,[2] AND GARY A. BANNON[2]

Department of Biology, Washington University, St. Louis, Missouri 63130,[1] and Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, Arkansas 72205[2]

Received 5 February 1990/Accepted 9 August 1990

The DNA sequences of a cDNA clone and the macronuclear genomic fragment corresponding to the functional copy of the *SerH3* surface antigen gene of *Tetrahymena thermophila* were determined. Primer extension and nuclease protection assays show that the *SerH3* transcription unit is 1,425 nucleotides long and contains no introns. The predicted polypeptide encoded by the *SerH3* gene has a molecular mass of 44,415 daltons; one-third of its 439 residues are either cysteine, serine, or threonine. The central half of the polypeptide consists of three homologous domains in tandem array; within these domains, the cysteine, proline, and tryptophan residues occur in highly regular patterns.

Each cell of the ciliated protozoan *Tetrahymena thermophila* is covered by one of five types of surface antigens that are encoded by five unlinked loci and expressed in a mutually exclusive, environmentally determined manner (for a review, see reference 29). Condition-dependent variation in surface antigens during vegetative growth has been observed not only in other ciliates such as *Paramecium* species (5, 8, 20) but also in very distantly related protozoans, such as trypanosomes (2). Such antigen switching has obvious biological significance for trypanosomes, which spend part of their life cycle as endoparasites of immunocompetent vertebrates. Its significance for free-living ciliates is far less obvious; however, it does provide a potentially powerful system for analyzing molecular mechanisms of gene regulation in these organisms.

Three of the five classes of *Tetrahymena* antigens are regulated by the environmental temperature: L is expressed when the cells are grown below 20°C, H is expressed between 20 and 35°C, and T is expressed above 35°C. Expression of two other antigen classes, S and I, is regulated by the composition of the growth medium (presence of 200 mM NaCl or low concentrations of anti-H antibodies, respectively) and overrides the expression of the temperature-regulated loci.

Molecular characterization of the *SerH3* allele of the H surface antigen class has been facilitated by the isolation of a cDNA clone, pC6, that was shown to hybridize to the *SerH3* mRNA (4, 18). pC6 hybridizes to a small family of related sequences in both the macro- and micronuclear DNA (15, 19, 30). The member of this sequence family encoding pC6 and thus, an active *SerH3* gene, was identified by using a subclone of pC6 (pGpC6-295) as a hybridization probe. pGpC6-295 hybridizes to one macronuclear DNA fragment from cells carrying the *SerH3* allele and does not hybridize to macronuclear DNA samples from cells homozygous for any of the other *SerH* alleles (15, 30).

Molecular analysis of *SerH3* expression has shown that SerH3 synthesis is controlled primarily at the posttranscriptional level. In vitro nuclear runoff transcription assays were used to demonstrate that the rate of synthesis of the *SerH3* mRNA remains essentially the same whether the cells are grown at 30 or at 40°C. However, whereas the RNA from cells grown at 30°C has a half-life of greater than 60 min, its half-life falls to ~3 min when the cells are grown at 40°C (17). This indicates that the temperature-dependent synthesis of SerH3 antigen is regulated predominantly at the level of mRNA stability. The mechanism(s) of this temperature-dependent difference in mRNA stability has not been elucidated.

In order to examine the SerH3 gene structure and expression, a genomic clone, λgt501, was isolated by screening a *Tetrahymena* macronuclear DNA library with ^{32}P-labeled pGpC6-295 insert. This clone contained a single *Eco*RI, *Hin*dIII, or *Bgl*II fragment that hybridized to pGpC6-295 or to pC6 and comigrated with a corresponding macronuclear DNA fragment (data not shown). This clone was mapped, and the regions of pC6 hybridization were determined (Fig. 1).

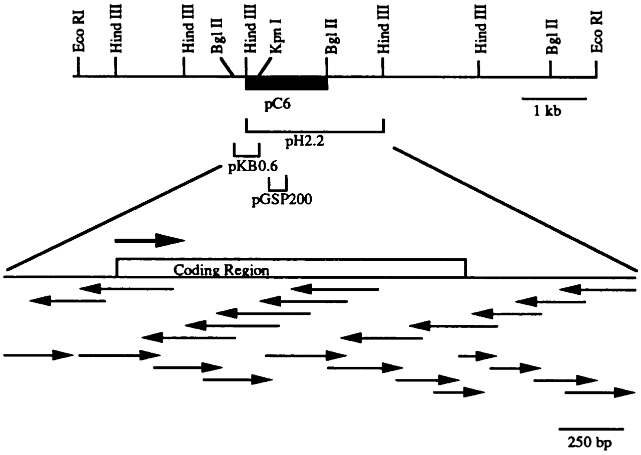

FIG. 1. Restriction map of the macronuclear genomic clone λgt501. The region of pC6 hybridization is indicated by the black rectangle. The bottom line shows an enlargement of the sequenced region with the coding region and direction of transcription depicted above it; the thin arrows indicate the sequencing strategy.

* Corresponding author.

† Present address: La Jolla Cancer Research Foundation, 10901 N. Torrey Pines Road, La Jolla, CA 92037.

‡ Present address: Chemistry Program, Southern Arkansas University, Magnolia, AR 71753.

```
-412                         GG  ATCTAAATTG  TTATTATAAT  ACTAAAATTG  ATTGCTTTGG  ATTACTTGTT
-360 TGCTAAAACT  AAACTTGATA  ATTCAATTCT  ATTGTCTGAT  ATTTTAAAAT  TTATTTTATA  TTTTTGAAAA  GTTACATCAA
-300 TTTTATGAAA  ATAAGATAAA  ATATTGTTTA  TTTCAAACTA  AAAATAACCA  AAATAATTAA  TTTTCTTTAT  TTTGAGTTAA
-240 AAAAAATTAG  TTAAAACATA  AACTAATACC  TCAGATTTAA  ATATACTTAA  AGGATATTTA  TTTCTTTTTA  ATTATTTAAA
-160 AATTAAATTT  TTCATCTATT  TTTAATTAAT  ATTTTGCAAG  CTTTTTCTTC  TTTTTAATAA  TTATTACTTT  CGTTTGTCTA
 -80 AATTTGTGTG  TTTATATCTT  TCTCGCTTTT  CTTTAACTAT  AATAATTAAT  TAATTCAAAC  AAAAAATTCA  AAAAAAAAAA

       M   Q   N   K   T   I   I   I   C   L   I   I   S   Q   L   L   V   S   V   F     20
    1 ATG TAA AAC AAA ACT ATA ATA ATT TGC TTA ATA ATT TCT TAA CTT CTG GTT TCT GTA TTT

       S   S   A   G   G   Q   A   N   C   T   G   V   A   A   G   T   D   C   A   S     40
   60 TCA TCA GCA GGA GGT CAA GCT AAT TGT ACA GGT GTT GCT GCT GGT ACC GAT TGT GCT AGT
                                                                      TGT GCT AGT

       V   C   G   V   P   T   V   A   G   T   G   T   T   A   C   S   W   V   S   S     60
  121 GTC TGT GGA GTA CCT ACA GTT GCA GGA ACT GGT ACA ACA GCT TGT AGT TGG GTT AGT TCT
      GTC TGT GGA GTA CCT ACA GTT GCA GCA ACT GGT ACA ACA GCT TGT AGT TGG GTT AGT TCT

       S   T   L   T   T   C   T   V   T   D   C   T   C   L   T   T   G   T   V   T     80
  181 TCT ACT TTG ACC ACT TGC ACT GTT ACT GAT TGT ACT TGC CTA ACT ACT GGT ACT GTA ACT
      TCT ACT TTG ACC ACT TGC ACT GTT ACT GAT TGT ACT TGC CTA ACT ACT GGT ACT GTA ACT

       G   I   T   N   L   N   D   Q   F   C   T   S   C   K   G   S   T   S   N   T    100
  241 GGT ATC ACT AAT TTA AAT GAT TAA TTT TGT ACT TCT TGT AAA GGA TCT ACC TCA AAT ACC
      GGT ATC ACT AAT TTA AAT GAT TAA TTT TGT ACT TCT TGT AAA GGA TCT ACC TCA AAT ACC

       Y   A   N   G   A   G   T   A   C   V   A   A   S   A   S   C   N   S   T   I    120
  301 TAT GCT AAT GGT GCT GGA ACT GCT TGT GTA GCT GCT TCT GCT TCA TGC AAC AGC ACC ATA
      TAT GCT AAT GGT GCT GGA ACT GCT TGT GTA GCT GCT TCT GCT TCA TGC AAC AGC ACC ATA

       R   G   T   T   A   W   T   V   G   D   C   T   V   C   T   P   T   T   P   A    140
  361 AGA GGA ACT ACT GCA TGG ACT GTT GGT GAT TGC ACC GTT TGT ACT CCT ACT ACC CCT GCA
      AGA GGA ACT ACT GCA TGG ACT GTT GGT GAT TGC ACC GTT TGT ACT CCT ACT ACC CCT GCA

       L   V   G   S   T   C   K   A   C   N   T   I   S   S   A   W   T   D   A   N    160
  421 TTG GTT GGT AGT ACT TGT AAG GCT TGT AAT ACT ATA AGT AGT GCA TGG ACT GAT GCA AAT
      TTG GTT GGT AGT ACT TGT AAG GCT TGT AAT ACT AT. ... ... ... ... ... ... ... ...

       C   A   A   C   A   S   T   S   T   P   K   G   N   T   N   F   A   N   S   A    180
  481 TGT GCT GCA TGC GCC AGT ACT TCT ACC CCT AAA GGT AAC ACA AAT TTT GCT AAC TCT GCT
      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

       G   T   A   C   V   N   A   S   A   T   C   A   S   G   S   R   G   T   T   A    200
  541 GGT ACT GCT TGT GTT AAT GCT TCC GCA ACA TGT GCT AGT GGT AGT AGA GGT ACT ACT GCT
      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

       A   N   A   W   T   V   A   D   C   L   A   C   T   P   A   T   P   V   F   V    220
  601 GCC AAT GCT TGG ACA GTT GCT GAT TGT CTT GCT TGT ACT CCT GCT ACT CCT GTT TTC GTA
      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

       P   A   A   S   P   A   V   T   T   S   C   V   A   C   S   A   A   T   S   G    240
  661 CCC GCT GCT TCC CCT GCA GTA ACT ACT TCT TGT GTT GCT TGC TCT GCT GCC ACT TCA GGT
      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

       L   N   D   A   L   C   N   A   C   A   S   S   A   S   P   A   A   K   T   T    260
  721 TTG AAT GAT GCC TTA TGT AAT GCT TGT GCA TCA AGT GCT TCC CCT GCA GCT AAA ACT ACT
      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

       F   A   N   T   A   G   S   A   C   V   A   S   S   A   T   C   T   A   G   S    280
  781 TTT GCT AAT ACT GCT GGT TCT GCT TGT GTT GCT TCT TCC GCA ACA TGC ACT GCT GGT AGT
      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
```

```
             R   G   T   T   A   A   N   A   W   T   A   A   D   C   L   A   C   T   P   A        300
 841        AGA GGT ACT ACT GCT GCC AAT GCT TGG ACA GCT GCT GAT TGT CTT GCT TGT ACA CCT GCT
            ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

             T   P   A   V   Q   F   G   A   S   P   A   T   T   S   S   C   V   A   C   N        320
 901        ACT CCT GCC GTA CAA TTT GGT GCT TCT CCT GCT ACT ACT TCT AGT TGT GTT GCT TGT AAT
            ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

             T   I   N   S   G   W   T   D   A   N   C   N   S   C   A   M   L   L   A   L        340
 961        ACT ATT AAT TCA GGA TGG ACA GAT GCT AAC TGT AAT TCA TGT GCT ATG CTG CTA GCC CTT
            ... ..T AAT TCA GGA TGG ACA GAT GCT AAC TGT AAT TCA TGT GCT ATG CTG CTA GCC CTT

             K   Q   K   I   S   S   L   R   L   M   E   V   L   V   Q   Q   L   C   F   H        360
1021        AAA CAA AAA ATA TCG TCG CTA AGG CTG ATG GAA GTG CTT GTG TAG CAG CTG TGT TTT CAT
            AAA CAA AAA ATA TCG TCG CTA AGG CTG ATG GAA GTG CTT GTG TAG CAG CTG TGT TTT CAT

             A   L   N   L   L   E   V   Q   I   N   G   L   M   Q   T   V   P   P   A   M        380
1081        GCA CTC AAT CTG CTA GAG GTT CAA ATA AAT GGA CTA ATG CAG ACT GTG CCG CCT GCA ATG
            GCA CTC AAT CTG CTA GAG GTT CAA ATA AAT GGA CTA ATG CAG ACT GTG CCG CCT GCA ATG

             V   L   L   L   M   Q   I   N   M   P   L   L   M   V   L   H   V   K   L   H        400
1141        GTA CTG CTG CTA ATG CAA ATC AAT ATG CCT CTG CTG ATG GTT CTA CAT GTC AAG CTA CAT
            GTA CTG CTG CTA ATG CAA ATC AAT ATG CCT CTG CTG ATG GTT CTA CAT GTC AAG CTA CAT

             R   L   L   V   L   S   V   V   R   S   L   L   A   F   Y   Q   F   Y   L   L        420
1201        AGG CTT CTA GTA CTT TCA GTG GTT AGA TCT TTG TTA GCA TTT TAT TAG TTT TAT CTG CTT
            AGG CTT CTA GTA CTT TCA GTG GTT AGA TCT TTG TTA GCA TTT TAT TAG T

             C   Q   F   D   Y   S   S   F   K   Q   K   F   R   I   Q   N   C   T   F            439
1261        TGT TAA TTT GAT TAT TCA AGT TTC AAA CAA AAA TTT AGA ATT TAA AAT TGC ACT TTT TGA

1320  TTGTTGGTAT  CTATATTTTT  ATGTGTACTG  ATTTGATTAA  AACATGTGTA  AAATTATCAT  TTTAATCTAG  ATAAGAATTT

1400  TAATTTCAAA  ATAAACTATT  TAACACAAAA  TTTATTAAAT  TTTAAAATAA  TAAATTTAGA  CTTTATTTTA  ATTATTATCA
1480  GAATATAACT  GGTATCCTTA  AAATTTGGAA  ATGAATTTAT  TTATTAAATG  CAAGTATTAA  TTTAAATAAA  AATTAATTCT
1560  AATACTAAAC  GAAAATAACA  CAAGATTTAA  AAAAGCAAAA  AGAAATAAAA  ATATTACAAA  AATGGAATAA  AATAAACAAT
1640  TTTTATATGC  CATTAAATTC  AAATAATATT  ACATTTATAA  TCATCTAAAT  AGTTTTTAGA  TTATTCTGAA  TACCTTACCA
1720  TGTCTGTCTT  CCATTCTAAA  AGTAAACTTT  TCAAAATTTG  TCTGAAATTA  AAAATGAAAC  ATCGATAACT  CCAAAATTTT
1800  TGAATTTGGG  GGTTGCGCGT  TAATGAGGTT  TTACAGTATT  TGCTTTATTA  CTCTTTAATT  ATCAGTGATC  AAACAATTTT
1880  AAAATAGCAA  ATTACTATGT  AAATATAATA  GATTTAAAAT  TTGTATTTTA  AAAAAATTAA  ATGATTACTT  ATTTCACTTA
1960  AAAGAAAAAT  ATACTCAATT  ACTAATAAAA  AATATTGAAG  AGGAAAGAAG  GCAACTGAAT  GATAATCATG  TTATTGTTTT
2040  CTTTCGCTAA  TTAATGAATT  ATAATAAAAA  ATTATTTTGG  TAGTAAATTG  ATCAATTGGT  TGAAGCTT
```

FIG. 2. Sequence of the *Tetrahymena SerH3* gene. The complete nucleotide sequence of pH2.2 plus pKB0.6 are shown on the first line. The sequence was determined by using a series of unidirectional subclones generated by exonuclease III digestion (12) and sequenced by the methods of Sanger et al. (26) as modified for double-stranded DNA (31). The sequence of pC6, determined by the dideoxy method using two pBR322/PstI primers (31), is aligned to the λgt501 genomic sequence and shown on the second line. The pC6 sequence is missing a short stretch of nucleotides closest to the primers. The underlined sequence 5′ to the coding region of λgt501 represents a potential TATA box, and the bold T's (at positions −38, −34, and −30) define the putative 5′ ends of the transcripts, as determined by primer extension. The bold C at position 1387 defines the 3′ end of the *SerH3* message, as determined by nuclease protection. The predicted amino acid sequence is represented by the single-letter code above the λgt501 sequence. The underlined amino acid residues (positions 3 to 5, 28 to 30, 117 to 119, 186 to 188, and 436 to 438) constitute glycosylation consensus sequences. The numbers on the left of the figure indicate the nucleotide sequence (with the first base of the translation initiation codon set at +1), and those on the right correspond to the predicted amino acid sequence.

We have sequenced the pC6 cross-hybridizing region of λgt501 and most of the original cDNA clone, pC6. Since preliminary sequencing of the pC6 clone did not reveal a poly(A) stretch [even though the cDNA synthesis was originally primed with oligo(dT)], we first sequenced the entire 2.2-kb *Hin*dIII fragment of λgt501 (pH2.2). The sequencing strategy is summarized in Fig. 1. The initial analysis of this sequence indicated that pH2.2 contained the entire coding sequence, but only a very short stretch of the 5′ untranscribed region. For this reason, the overlapping 0.6-kb *Kpn*I-*Bgl*II fragment, pKB0.6, was also sequenced (Fig. 1). Altogether, 2,558 nucleotides of the genomic clone λgt501 were sequenced on both strands (Fig. 2); the sequence is less than 25% GC in the first 454 and the last 780 nucleotides, while the central region has a GC content of approximately 50%.

cDNA clone pC6 was also sequenced by the dideoxy method using two pBR322/PstI primers (31). The pC6 sequence has been aligned with the genomic sequence; for simplicity, all nucleotides are numbered with reference to the predicted translation start site of the λgt501 sequence (Fig. 2). The pC6 sequence aligns well with the λgt501 sequence, except that it possesses a 510-bp gap in the region corresponding to nucleotides +456 to +965 of the genomic

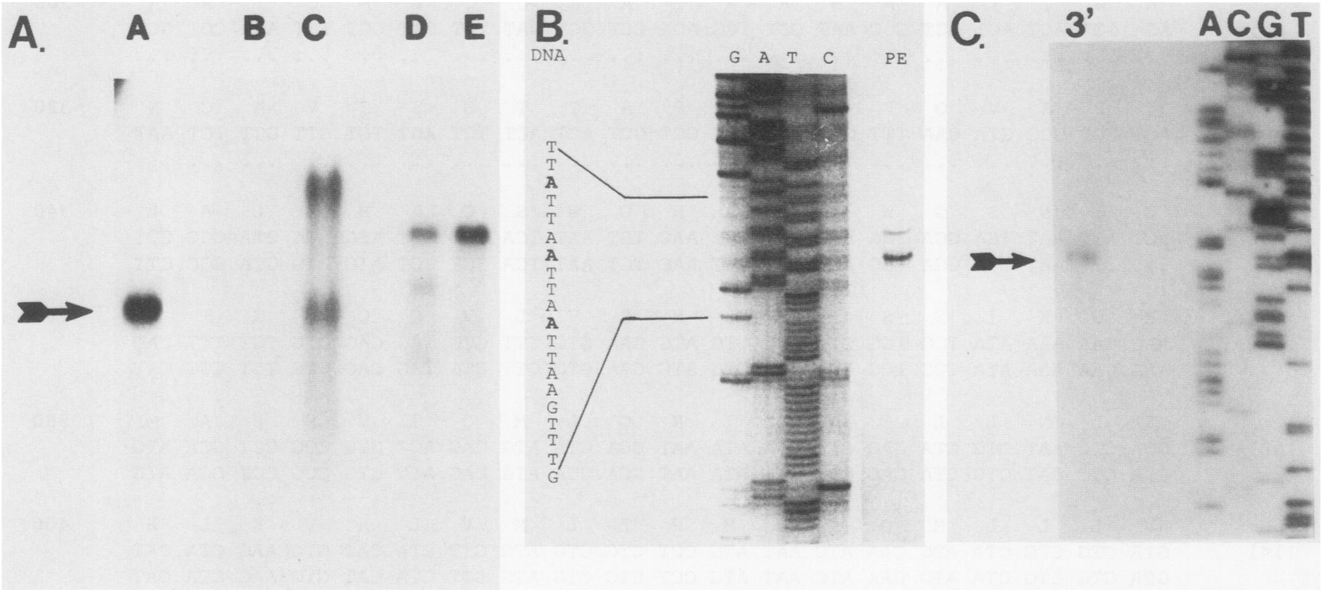

FIG. 3. Mapping the *SerH3* mRNA. (A) Asymmetrically transcribed RNAs were prepared from clone pH2.2 as described in the text. Total cytoplasmic RNA and ^{32}P-labeled transcripts were subjected to electrophoresis on 1.0% formaldehyde agarose gels and transferred to nitrocellulose. Lane A, Total cytoplasmic RNA. This lane was cut from the blot and hybridized with ^{32}P-labeled pC6 to locate the full-length *SerH3* mRNA. Lane B, RNA representing the DNA sense strand followed by the solution hybridization with *Tetrahymena* RNA and RNase digestion. No RNAs were protected from nuclease digestion. Lane C, RNA representing the DNA antisense strand followed by solution hybridization with *Tetrahymena* RNA and RNase digestion. An RNA which comigrated with the intact *SerH3* mRNA (lane A) but smaller than the undigested ^{32}P-labeled transcript (lane E) was protected. Lane D, Full-length RNA transcript representing the DNA sense strand. Lane E, Full-length RNA transcript representing the DNA antisense strand. Arrow indicates full-length cytoplasmic SerH3 mRNA. (B) Mapping the 5′ end of the SerH3 mRNA. Autoradiograph of a DNA sequencing gel. Lanes G, A, T, and C, Sequencing reactions using clone pH2.2 (contains the 5′ end of the *SerH3* gene) and a 5′-end-labeled 57-nucleotide *Asp* 718 I-*Alu*I fragment (nucleotides +102 to +158) from this clone as primer. Lane PE, Products of a primer extension experiment using the *Asp* 718 I-*Alu*I fragment and 10 μg of *Tetrahymena* poly(A)$^+$ RNA. Bold A's represent the 5′ ends of the *SerH3* mRNA; the band corresponding to the third A becomes apparent only after longer exposure of the gel. (C) Mapping the 3′ end of the *SerH3* mRNA. Autoradiograph of a DNA sequencing gel. The same clone and procedure as described in panel A were used except that the clone was linearized with *Bgl*II (located 91 bp upstream from the protein synthesis termination codon). Protected fragments were electrophoresed on an 8% DNA sequencing gel (lane 3′). Dideoxy sequencing reactions of a pGEM (Promega) were used as molecular size markers (lanes A, C, G, and T). Arrow indicates a protected band of 158 ± 2 nucleotides.

sequence. The exact correspondence between these two sequences—apart from the one gap—confirms the conclusion drawn from pGpC6-295 hybridization: namely, that λgt501 contains the gene encoding pC6, and hence a functional *SerH3* gene.

The portion of the λgt501 sequence that is not present in pC6 represents a deletion in pC6 that occurred during cloning or propagation in *Escherichia coli*. This conclusion was demonstrated by using pGSP200, a 192-bp *Sph*I-*Pst*I fragment of pH2.2 that spans nucleotides +618 to +799 of λgt501, to probe a Northern blot of RNA from *Tetrahymena* cells that had been grown at either 30 or 40°C. pGSP200 and pC6 hybridized to the same temperature-regulated RNA species (data not shown). The transcript-mapping data described below further indicate that the 510-bp deletion found in pC6 is a cloning artifact and not a reflection of an intron.

As mentioned above, *SerH3* mRNA stability plays a major role in regulating *SerH3* expression. Changes in RNA stability have been shown to be important in the regulation of a number of other genes including β-tubulin, c-*fos*, and the lymphokine granulocyte macrophage colony-stimulating factor (7, 27, 28). In the case of granulocyte macrophage colony-stimulating factor, an AUUUA sequence repeat in the 3′ untranslated region of the mRNA has been shown to be the *cis*-acting element involved in determining the stability of the message under different conditions (28). The boundaries of the *SerH3* mRNA were mapped to the λgt501 genomic clone to determine whether this *cis*-acting element was present within the mature transcript.

The pH2.2 insert was subcloned into a pGEM vector and used to transcribe strand-specific RNA probes with the T7 RNA polymerase as described by Horowitz et al. (13). Each of the probes was then hybridized to total RNA from cells expressing SerH3 antigen and digested with ribonuclease A and T$_1$, and the protected fragments were electrophoresed on a formaldehyde-agarose gel. The results show the presence of a nuclease-protected fragment that is the same size as the *SerH3* mRNA, indicating that the *SerH3* gene does not contain any introns (Fig. 3A). The largest protected band in lane C of Fig. 3A was reproducible from experiment to experiment; however, this band cannot represent an intron since it is migrating as a larger molecule than the undigested ^{32}P-labeled transcript (Fig. 3A, lane E).

In order to map the 5′ end of the SerH3 transcript, a 57-nucleotide primer corresponding to the *Asp* 718 I-*Alu*I region of pH2.2 (nucleotides +102 to +158) was gel purified and used to prime synthesis of a DNA strand complementary to the message. The same primer was used to sequence pH2.2. All of the reaction products were electrophoresed on a 6% sequencing gel and autoradiographed. Two strong transcription initiation sites were detected at −30 and −34 relative to the A of the ATG translation initiation codon. A third, but weaker, transcription start site at −38 becomes apparent after longer exposure of the gel (Fig. 3B). A

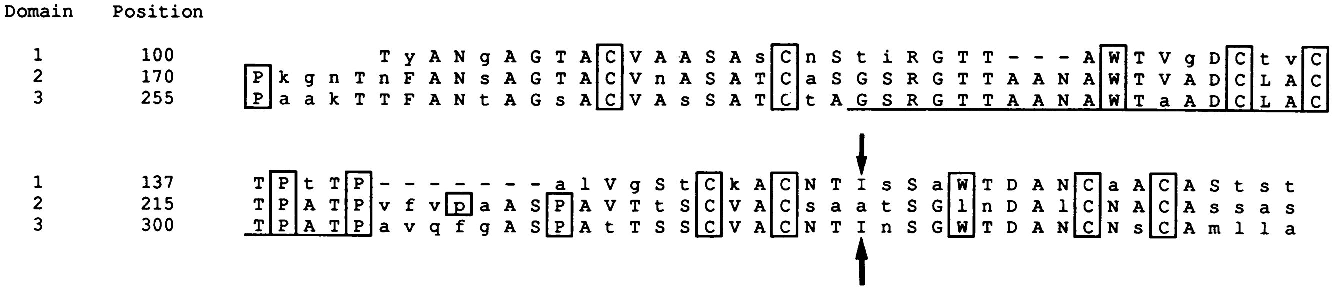
Domain Position
1 100 T y A N g A G T A C V A A S A s C n S t i R G T T - - - A W T V g D C t v C
2 170 P k g n T n F A N s A G T A C V n A S A T C a S G S R G T T A A N A W T V A D C L A C
3 255 P a a k T T F A N t A G s A C V A s S A T C t A G S R G T T A A N A W T a A D C L A C
1 137 T P t T P - - - - - - - a l V g S t C k A C N T I s S a W T D A N C a A C A S t s t
2 215 T P A T P v f v p a A S P A V T t S C V A C s a a t S G l n D A l C N A C A s s a s
3 300 T P A T P a v q f g A S P A t T S S C V A C N T I n S G W T D A N C N s C A m l l a

FIG. 4. Tandemly repeated domains in the *SerH3* sequence. Residues 100 to 339 in the predicted polypeptide sequence have been aligned to reveal the degree of similarity between three successive domains. Gaps required to give optimal alignment of the first domain with the other two are indicated by dashes. Capital letters are used for positions at which two or three of the domains have the same amino acid, and lowercase letters are used for unique residues. The cysteine, proline and tryptophan resides are boxed to emphasize alignments. The arrows indicate the limits of the region of *SerH3* that was deleted in pC6, and the underscoring indicates the region of highest homology.

potential TATA box is present between nucleotides −64 and −69 (Fig. 2).

The 3′ end of the SerH3 transcription unit was determined by linearizing pH2.2 at a *Bgl*II site located 91 nucleotides upstream of the TGA translation termination codon and performing an RNase protection experiment utilizing total *Tetrahymena* cytoplasmic RNA or poly(A)$^+$ RNA as described above. The RNase-resistant fragments were electrophoresed on an 8% sequencing gel and autoradiographed. The data obtained indicate that the 3′ end of the *SerH3* mRNA is located 67 ± 2 nucleotides downstream of the TGA translation termination codon (Fig. 3C).

The *cis*-acting stability sequence, AUUUA, was found three times within the coding region of the mature *SerH3* mRNA transcript (beginning with nucleotides 251, 1293, and 1300), but not in the 3′ untranslated portion of the mature transcript. If this sequence appeared randomly in *Tetrahymena* DNA (75% AT; 9), it would appear seven times every 1,000 bp. However, since the *SerH3* coding region is ~50% GC, this consensus sequence appears at about the correct frequency for randomness (once every 1,000 bp). If the higher AT content of the transcribed but untranslated portions of *Tetrahymena* mRNAs were taken into account, one would expect the sequence AUUUA to occur ~18 times every 1,000 bp. Interestingly, this portion of the *SerH3* transcript does not contain the consensus sequence. Whether this sequence plays the same role in *Tetrahymena* surface protein mRNA stability as it does in other systems remains to be experimentally determined.

Because coding sequences impose constraints on nucleotide composition, their presence in AT-rich genomes such as that of *Tetrahymena* species (75% AT; 9) can be readily detected due to their higher GC content than the genome in general (3, 21). Furthermore, the genetic code used by ciliates differs from the universal code in that the triplets UAA and UAG are not always used as translation stop codons; and in the case of *Tetrahymena* species, these triplets code for glutamine or glutamic acid (10, 11, 14, 16). Considering these characteristics and the mRNA mapping data discussed above, computerized examination of the entire sequenced region of λgt501 revealed only one open reading frame of significant length. It is 1,320 nucleotides long, begins with an ATG initiation codon, ends with a TGA termination codon, contains five TAA and two TAG codons, and codes for a predicted polypeptide of 439 amino acids in length, with a molecular mass of 44,415 Da. The deduced protein sequence (Fig. 2) contains five asparagine-linked glycosylation consensus sequences, has an N-terminal region that is characteristic of a signal peptide, is relatively cysteine rich (37 residues), and includes either cysteine, serine, or threonine as one-third of its residues.

The most striking feature of the deduced amino acid sequence of λgt501 is that the central half of the polypeptide (amino acid residues 100 to 339) consists of three highly homologous sequence domains in tandem array (Fig. 4). The second and third of these domains are both 85 residues long, have 65% sequence identity overall, and differ in only one amino acid residue out of 24 in a region (underscored in Fig. 4) near the center of each domain. Domain 1 is sufficiently divergent that to align it with domains 2 and 3 required two short deletions; yet (deletions aside) it exhibits about 60% sequence identity with each of the other two domains. Perhaps the strongest indication that these three regions are truly homologous domains is that when they are aligned in this fashion, all 24 of the cysteine residues, all 5 of the tryptophan residues, and 10 of the 11 proline residues that they contain fall into register (these residues are boxed in Fig. 4).

A search of the GenBank database did not reveal any other sequences with significant homology to the *SerH3* sequence. However, by virtue of its internally repetitive structure, the *Tetrahymena* SerH3 antigen resembles several other protozoan surface antigens that have been analyzed previously. For example, the circumsporozoite protein of six *Plasmodium* strains, a parasitic protozoan responsible for causing malaria, has central domains of tandemly repeated amino acids (6, 22). Repetitive elements are also present in surface proteins of the parasites *Toxoplasma gondii* (25), *Trypanosoma cruzi* (23), and *Schistosoma mansoni* (1). This type of surface protein structure is not unique to parasitic protozoa. In a free-living ciliated protozoan, *Paramecium primaurelia*, the G surface protein (expressed when the cell is incubated between the temperatures of 14 and 32°C) contains a 74-amino-acid sequence that is repeated five times in tandem (24). It is interesting to note that even though the amino acid sequences of these surface proteins are quite dissimilar, they all show the same repeated nature in the middle third of the protein. It is possible that this conserved structure may imply conservation of function. The function of environmentally or developmentally regulated surface proteins remains to be determined.

M.M.T. would like to thank David and Marilyn Kirk for providing the stimulating and supportive environment in which this work was conducted. We thank D. Kirk for critical review of the manuscript and Pamela J. McMillan for excellent technical assistance.

M.M.T. was supported in part by the Department of Biology at

Washington University and Public Health Service grant GM-27215 from the National Institutes of Health to David Kirk. Other support came from National Science Foundation grant DMB-8716407 to G.A.B. and a graduate student research grant to H.D.L. from the University of Arkansas for Medical Sciences. R.L.W. was supported by a Research Opportunity Award from the National Science Foundation and a Southern Arkansas University Research Fund grant.

LITERATURE CITED

1. **Bobek, L., D. M. Rekosh, H. VanKeulen, and P. T. LoVerde.** 1986. Characterization of a female-specific cDNA derived from a developmentally regulated mRNA in the human blood fluke *Schistosoma mansoni*. Proc. Natl. Acad. Sci. USA **83:**5544–5548.
2. **Borst, P., and G. A. M. Cross.** 1982. Molecular basis of trypanosome antigenic variation. Cell **29:**291–303.
3. **Cupples, C. G., and R. E. Pearlman.** 1986. Isolation and characterization of the actin gene from *Tetrahymena thermophila*. Proc. Natl. Acad. Sci. USA **83:**5160–5164.
4. **Doerder, F. P., and R. L. Hallberg.** 1989. Identification of a cDNA coding for the SerH3 surface protein of *Tetrahymena thermophila*. J. Protozool. **36:**304–307.
5. **Forney, J. D., L. M. Epstein, L. B. Preer, B. M. Rudman, D. J. Midmayer, W. H. Klein, and J. R. Preer, Jr.** 1983. Structure and expression of genes for surface proteins in *Paramecium*. Mol. Cell. Biol. **3:**466–474.
6. **Galinski, M. R., D. E. Arnot, A. H. Cochrane, J. W. Barnwell, R. S. Nussenzweig, and V. Enea.** 1987. The circumsporozoite gene of the *Plasmodium cynomolgi* complex. Cell **48:**311–319.
7. **Gay, D. A., T. J. Yen, J. T. Y. Lau, and D. W. Cleveland.** 1987. Sequences that confer β-tubulin autoregulation through modulated mRNA stability reside within exon 1 of a β-tubulin mRNA. Cell **50:**671–679.
8. **Godiska, R.** 1987. Structure and expression of the H surface protein gene of *Paramecium* and comparison with related genes. Mol. Gen. Genet. **208:**529–536.
9. **Gorovsky, M. A.** 1980. Genome organization and reorganization in *Tetrahymena*. Annu. Rev. Genet. **14:**203–239.
10. **Hanyu, N., Y. Kuchino, and S. Nishimura.** 1986. Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* tRNAs. EMBO J. **5:**1307–1311.
11. **Harper, D. S., and C. L. Jahn.** 1989. Differential use of termination codons in ciliated protozoa. Proc. Natl. Acad. Sci. USA **86:**3252–3256.
12. **Henikoff, S.** 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28:**351–359.
13. **Horowitz, S., J. K. Bowen, G. A. Bannon, and M. A. Gorovsky.** 1987. Unusual features of transcribed and translated regions of the histone H4 gene family of *Tetrahymena thermophila*. Nucleic Acids Res. **15:**141–160.
14. **Horowitz, S., and M. A. Gorovsky.** 1985. An unusual genetic code in nuclear genes of *Tetrahymena*. Proc. Natl. Acad. Sci. USA **82:**2452–2455.
15. **Kile, J. P., H. D. Love, Jr., C. A. Hubach, and G. A. Bannon.** 1988. Reproducible and variable rearrangement of a *Tetrahymena thermophila* surface protein gene family occur during macronuclear development. Mol. Cell. Biol. **8:**5043–5046.
16. **Kuchino, Y., N. Hanyu, F. Tashiro, and S. Nishimura.** 1985. *Tetrahymena thermophila* glutamine tRNA gene that corresponds to UAA termination codon. Proc. Natl. Acad. Sci. USA **82:**4758–4762.
17. **Love, H. D., Jr., A. Allen-Nash, Q. Zhao, and G. A. Bannon.** 1988. mRNA stability plays a major role in regulating the temperature-specific expression of a *Tetrahymena thermophila* surface protein. Mol. Cell. Biol. **8:**427–432.
18. **Martindale, D. W., and P. J. Bruns.** 1983. Cloning of abundant mRNA species present during conjugation of *Tetrahymena thermophila*: identification of mRNA species present exclusively during meiosis. Mol. Cell. Biol. **3:**1857–1865.
19. **Martindale, D. W., H. M. Martindale, and P. J. Bruns.** 1986. *Tetrahymena* conjugation induced genes: structure and organization in macro- and micronuclei. Nucleic Acids Res. **14:**1341–1354.
20. **Meyer, E., F. Caron, and A. Baroin.** 1985. Macronuclear structure of the G surface antigen gene of *Paramecium primaurelia* and direct expression of its repeated epitopes in *Escherichia coli*. Mol. Cell. Biol. **5:**2414–2422.
21. **Neilsen, H., P. H. Andearsen, H. Dresig, K. Kristiansen, and J. Engberg.** 1986. An intron in a ribosomal protein gene from *Tetrahymena*. EMBO J. **5:**2711–2717.
22. **Ozaki, L. S., P. Svec, R. S. Nussenzweig, V. Nussenzweig, and G. N. Godson.** 1983. Structure of the *Plasmodium knowlesi* gene coding for the circumsporozoite protein. Cell **34:**815–822.
23. **Peterson, D. S., R. A. Wrightsman, and J. E. Manning.** 1986. Cloning of a major surface antigen gene of *Trypanosoma cruzi* and identification of a nanopeptide repeat. Nature (London) **322:**566–568.
24. **Prat, A., M. Katinka, F. Caron, and E. Meyer.** 1986. Nucleotide sequence of the *Paramecium primaurelia* G surface protein. J. Mol. Biol. **189:**47–60.
25. **Rodriguez, C., D. Afchain, A. Capron, C. Dissous, and F. Santro.** 1985. Major surface protein of *Toxoplasm gondii* (p30) contains an immunodominant region with repetitive epitopes. Eur. J. Immunol. **15:**747–749.
26. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
27. **Schuler, G. D., and M. D. Cole.** 1988. GM-CSF and oncogene mRNA are independently regulated in trans in a mouse monocytic tumor. Cell **55:**1115–1122.
28. **Shaw, G., and R. Kamen.** 1986. A conserved Au sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation. Cell **46:**659–667.
29. **Sonneborn, T. M.** 1974. *Tetrahymena pyriformis*. Handb. Genet. **2:**433–4676.
30. **Tondravi, M. M.** 1989. DNA rearrangements associated with the H3 surface antigen gene of *Tetrahymena thermophila* that occur during macronuclear development. Curr. Genet. **14:**617–626.
31. **Wallace, B. R., M. J. Johnson, S. V. Suggs, K. Miyoshi, R. Baht, and K. Itakura.** 1981. A set of synthetic oligodeoxynucleotide primers for DNA sequencing in plasmid vector pBR322. Gene **16:**21–26.