Evidence for a class of very small introns in the gene for hypoxanthine-guanine phosphoribosyltransferase in *Schistosoma mansoni*

Sydney P.Craig III, M.G.Muralidhar, James H.McKerrow[1] and Ching C.Wang

Department of Pharmaceutical Chemistry and [1]Department of Pathology, University of California, San Francisco, CA 94143, USA

Received October 31, 1988; Revised and Accepted December 21, 1988 EMBL accession nos*

ABSTRACT

The single copy gene for the hypoxanthine-guanine phosphoribosyltransferase (HGPRTase) of the parasitic trematode, *Schistosoma mansoni*, contains seven introns, the first four of which are only 31, 33, 42, and 32 bases in length. These are the smallest introns ever discovered in a non-viral nuclear gene coding for protein. These very small introns possess the canonical GT...AG splice site sequences but lack the branching sequence, the secondary structure, and the minimum size of approximately 50 bases believed to be required for the splicing of eucaryotic mRNA precursors. Evidently, a somewhat different splicing mechanism for the transcripts of these very small introns is necessary. Their discovery within the genes of helminths raises theoretical considerations for the evolution of introns in eucaryotes.

INTRODUCTION

HGPRTase has been a focus of interest in several different fields of biology and medicine because of its central role in the salvage of purines. Deficiency of HGPRTase has been implicated in gout (1) and Lesch-Nyhan syndrome (2) and is useful in the selection of mouse cell hybridomas (3). Furthermore, since many parasites are known to lack the enzymes for *de novo* biosynthesis of purine nucleotides (4), enzymes involved in the salvage of purine bases and nucleosides have been proposed as potential targets for antiparasitic chemotherapy (4). HGPRTase is of particular interest because it plays a major role in replenishing purine nucleotides for a number of parasites including *S. mansoni* (4).

Schistosomes are primitive metazoans and little is known about the structure of their genes. In an effort to determine details of the structure of the HGPRTase, we cloned and sequenced the cDNA encoding this enzyme in *S. mansoni* (4), one of three major species of trematode parasites that cause schistosomiasis (5). Subsequently, the cDNA was used to probe a bacteriophage library of genomic DNA from *S. mansoni* in order to permit detailed analysis of the gene(s) for HGPRTase. At the time, we had no reason to expect that organization of schistosomal genes would differ significantly from the organization of genes in other metazoa. However, analysis of the cloned gene revealed the presence of several introns smaller than any previously encountered within a eucaryotic nuclear gene coding for protein. The presence of the very small introns (vsi's) within an expressed copy of a gene raises important questions about current dogma of the physical constraints for removing transcripts of introns from mRNA precursors (6,7).

Within the last decade, many eucaryotic genes and their introns have been analyzed (see 8,9). The conservation of positions of introns within the same genes of different eucaryotes suggest that the introns are evolutionarily conserved and descended from the genes of common ancestors (10-12). Analysis of the intron structures of many genes and the splicing mechanisms employed by cells to remove transcripts of the introns from within nuclear mRNA precursors have led to a number of conclusions about the structural requirements for an intron (6,7). For example, the 5' and 3' limits of introns are almost always found to begin and end with the bases GT and AG, respectively (8,9). Current models for splicing transcripts of introns from nuclear mRNA precursors usually consist of two steps with the first resulting in the formation of the "lariat" structure between its 5' end and a branch site, typically 20-50 bp upstream of the 3' splice site (6,13). In higher eucaryotes, branching usually occurs at an adenine base just upstream of a pyrimidine rich sequence of 11 or more bases (6,9,14). Splicing is catalyzed by small nuclear ribonucleo-proteins (SnRNP's; 6,7,14), and the entire complex of DNA and SnRNP's, termed a splicesome, is believed to require a physical length of at least 50-55 bp for an intron transcript (6,15,16). Intron transcripts shorter than 50 bases can not be removed in theory by this mechanism of splicing, and efforts to splice artificially shortened introns from the yeast actin mRNA (16) or from rabbit β globin mRNA (15) have failed, although an artificial 29 bp intron has been shown to be spliced at a low level from an RNA polymerase B transcription unit within HELA cells (17).

MATERIALS AND METHODS

Mapping the schistosomal HGPRTase gene

Three bacteriophage clones containing genomic fragments from *S. mansoni* were identified and purified to homogeneity using radiolabeled cDNA encoding the schistosomal HGPRTase (4) as the probe to screen a λEMBL3 library (kindly provided by Geo. Newport of UCSF) containing fragments of genomic DNA from *S. mansoni.* Phage DNA was prepared as described elsewhere (4) except that the DNA was purified using the "glassmilk" procedure (Bio 101, Inc.) prior to removing the inserted schistosome DNA. The 12.8 kb insert of one clone, designated Smg1, was removed from the bacteriophage arms by cutting with the restriction endonuclease, *Sal* I, generating fragments 5.2 and 7.6 kb in length that were subsequently subcloned into

"Bluescript" plasmids (Stratagene Inc.). The DNA's from these subclones were subsequently digested with a variety of restriction endonucleases and the fragments were resolved in 1% agarose gels and analyzed according to the "Southern" procedure (18) using cDNA encoding the HGPRTase of *S. mansoni* as a probe. This analysis yielded information for the construction of a restriction map and indicated the general locations for sequences with homology to the HGPRTase cDNA. The exact positions on the map for the coding segments, or exons, was ultimately determined by correlating the data from restriction endonuclease analysis with the results from DNA sequence analysis.

Southern blot analysis

S. mansoni genomic DNA (Sm.) and purified plasmid DNA from the 5.2 and 7.6 kb subclones of the λEMBL3 insert of clone Smg1 (see Fig. 1) were cut with *Sal* I and *Eco* RI (S/E), *Bgl* II (Bg), *Hind* III and *Bgl* II (H/Bg), *Sal* I and *Xho* I (S/X), or with *Eco* RI and *Eco* RV (E/RV). Subsequently the samples were resolved in 1.0% Agarose - Tris/EDTA/borate gels before transferring to nitrocellulose as described (18). The filter was probed with a nick translated cloned partial cDNA (Smc1) for *S. mansoni* HGPRTase (4).

Sequence analysis

DNA sequences within the subcloned 5.2 and 7.6 Kb fragments of genomic DNA were determined using commercial primers from Stratagene and synthetic oligonucleotide primers from the Biomolecular Resource Center at UCSF, and a "Sequenase" kit from U.S. Biochemical Corporation following procedures outlined before (4). In order to analyze sequences for possible secondary structures within RNA transcripts, the program used was "rnafold" written by H. Martinez (19).

Materials

Restriction endonucleases were from Bethesda Research Laboratories. T_4 DNA ligase was from New England Biolabs. DNAase I and Proteinase K were from Sigma. Nitrocellulose was from Schleicher and Schuell. Radiolabelled nucleotides (^{32}P-dCTP and a-^{35}S-dATP) were from Amersham Corporation.

RESULTS

Restriction analyses

In order to isolate the gene encoding the HGPRTase of *S. mansoni*, we used cDNA encoding this enzyme (4) to probe a bacteriophage library of genomic DNA. Three clones were isolated and one, designated Smg1, was mapped with restriction enzymes (Fig. 1). In order to verify that the Smg1 clone is an accurate representation of the gene for HGPRTase within the genome of *S. mansoni*, we used nick translated cDNA to probe a Southern blot of restriction digested genomic DNA resolved in paral-

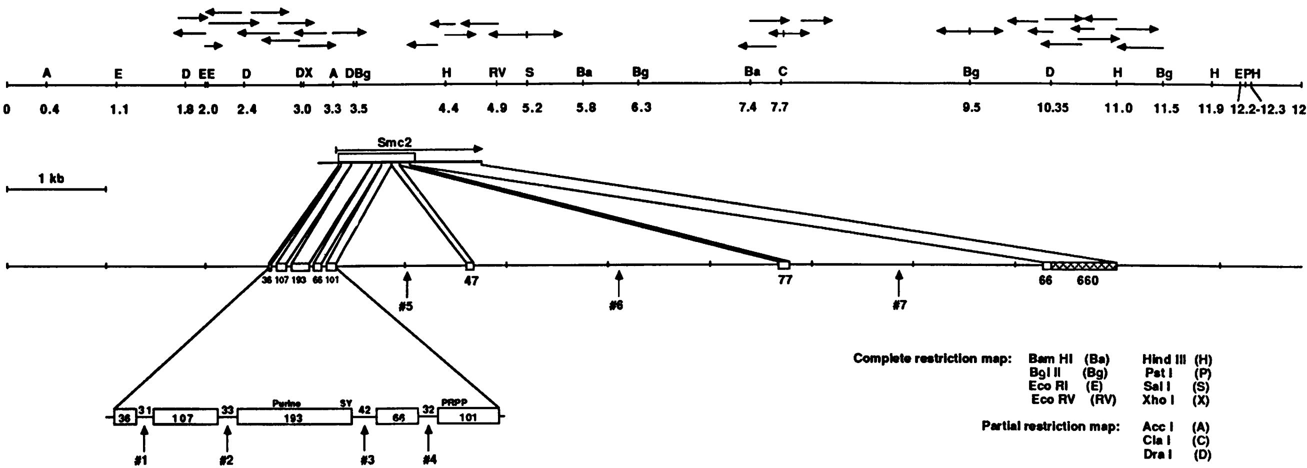

Figure 1. Map of the HGPRTase gene in *S. mansoni*. The arrow labeled Smc2 shows the length for the cDNA clone for the schistosomal HGPRTase (4). The lower of the two long horizontal lines shown represents the cloned 12.8 kb segment of *S. mansoni* genomic DNA. The upper long horizontal line shows the mapped positions for restriction endonuclease sites. The open boxed areas represent regions of the cDNA and gene that possess sequences coding for the enzyme. The transcribed cross-hatched area at the far right represents the long, probably untranslated, 3' tail for the mRNA. The number of bp found in each of the coding segments or exons and in each of the first 4 introns is shown. The arrows above the restriction map show the segments and directions for which nucleotide sequences have been determined.

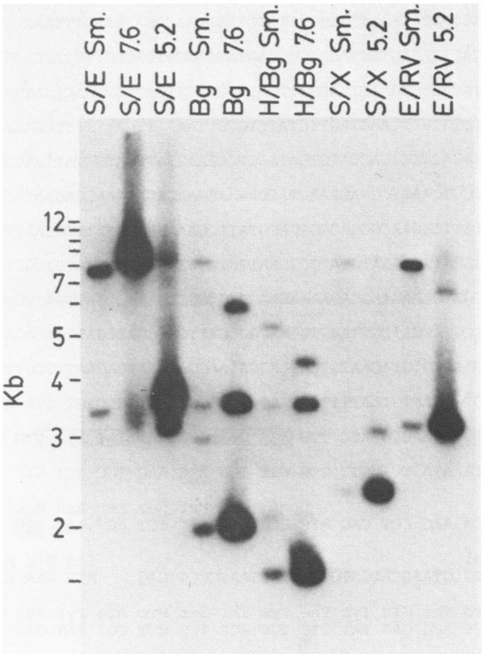

Figure 2. Southern blot of restriction endonuclease digested *S. mansoni* genomic DNA resolved in parallel with similarly digested cloned *S. mansoni* genomic DNA. *S. mansoni* genomic DNA (Sm.) and purified plasmid DNA from the 5.2 and 7.6 kb subclones of the 12.8 kb lambda EMBL3 insert of clone Smg1 (see Fig. 1) were cut with *Sal* I and *Eco* RI (S/E), *Bgl* II (Bg), *Hind* III and *Bgl* II (H/Bg), *Sal* I and *Xho* I (S/X), or with *Eco* RI and *Eco* RV (E/RV).

lel with similarly digested aliquots of subcloned fragments of the Smg1 clone (Fig. 2). The sizes for genomic fragments agree well with those predicted from the restriction map in Fig. 1, and most of the restriction fragments of genomic DNA correspond with restriction fragments of the subcloned gene. The remaining genomic fragments were from unmapped restriction sites located beyond the limits of the Smg1 clone. Thus, there appear to have been no major rearrangements during cloning, and Smg1 represents a clone of the sole copy of the HGPRTase gene within the genome of *S. mansoni.*

Sequence analyses

The exact locations of exon-intron boundaries were determined by sequencing the cloned genomic DNA (Fig. 3). Sequences thus determined yielded the locations for restriction endonuclease cleavage sites in or near the exons. This information was integrated into the restriction map of Smg1 to give the specific locations for the exons within the overall map (Fig. 1). The results show that the gene for HGPRTase in *S.*

```
CTTTGATTATTGTACTAATCTAATTGTTTTTTTGTCTAGAAAACAGATAAAGTTTAAAGTTGACCTTGTTA

TGACTATAGATTTCACATAGACTCTGAGATCGCCAGTATCGCGTTTACAGTTTTAGTAGTGTATCTGTAGG

AAAAACTAAATCCTACAGTGTGAACTAAGGATTCATTTGTACTTGATTGTTCACAGATTATGAATTCCAAA

TACAATACGTTACTTTGTGTTCAACTAGTTCTATCTGGCTTAAAATTCAATTTTTCGAGAATTCCTAGCTA

GTAAATTGATTCGAAGACTTCTGACCATCTCATAGCTCCAAAGTATCCATTGTATTATGTGTGTTTTCCTT

GACGTATAGAGGTCCCTCAAGTCCAACAAACTACCAGAAGGAAGAGAAAAACGAGAATAGGCAACATTCTC

TAATTATAGTATTTATTTAGAACGCGACGGTGACGTATTCCAAAGCACGAATTTGCAATTCGGATCGTCCT

AGGAATTTCGTATGACGTTTAAAATGTATTCGCAATCGTTTTTTGAGAAATGATGTCGCTAACCAAATGAA

TTGTCAAATGCTGCGATCAAAGAAGCAGAAAACAAGGCGTGGTAGAAACCGATGTATTCGAACTCACGAGT

AGCAGCATTAAGTTCGTATCATCGTTAAGGGAAACGCGCATTGTACGATCGATTGACACATACATGACGTC

GAATATCAATACAGAAGCGTGTACAATGTCACAACGATAAGGAAAAATCACAATTTCGTTTGCTCGTGTAA

AATGCCTGATTTTCTGTTATTGTAATTTTCGACATCA  ATG TTA ACT AGT TTA ATA ACA AGT

AGT ACG ACA GTT ACG CTG ACC TTG [AGC CAA] ATA TAT [TAT ATA T]TG GAC ATT GCT

TGT GGA TTC TTA ATA TCT GTT TTG GTT TGG ATG AAC TCC TCT GTC CTT GAT AAC

                                        Met Ser Ser Asn Met Ile Lys Ala
GGC AAC CAT TCA AAC CCT CAG ATC CGC GAC ATG TCT AGT AAC ATG ATA AAA GCT

Asp Cys Val Val                                    Ile Glu Asp Ser Phe
GAC TGT GTT GTG GTAAGCATGGTCTTTTTACTAATCCCTTCAG    ATA GAA GAC AGT TTT

Arg Gly Phe Pro Thr Glu Tyr Phe Cys Thr Ser Pro Arg Tyr Asp Glu Cys Leu
CGA GGA TTT CCT ACG GAG TAT TTC TGC ACA TCT CCT CGG TAT GAC GAA TGC TTG

Asp Tyr Val Leu Ile Pro Asn Gly Met Ile Lys Asp Ar
GAT TAT GTT CTC ATA CCA AAT GGT ATG ATA AAA GAT AGG TAGATCATAAATTCACGAT

             g Leu Glu Lys Met Ser Met Asp Ile Val Asp Tyr Tyr Glu Ala
AAGTATTTCAAAGG CTT GAA AAA ATG TCA ATG GAT ATT GTT GAC TAT TAC GAG GCC

Cys Asn Ala Thr Ser Ile Thr Leu Met Cys Val Leu Lys Gly Gly Phe Lys Phe
TGT AAT GCG ACA TCG ATC ACA CTT ATG TGT GTC CTC AAA GGT GGA TTT AAA TTC

Leu Ala Asp Leu Val Asp Gly Leu Glu Arg Thr Val Arg Ala Arg Gly Ile Val
CTT GCT GAT CTT GTT GAT GGG CTT GAA CGC ACT GTC CGT GCT CGA GGT ATC GTC

Leu Pro Met Ser Val Glu Phe Val Arg Val Lys Ser Tyr Val
CTA CCA ATG TCC GTT GAG TTT GTT CGT GTC AAG AGT TAT GTT GTACGTTAATAAATA

                               Asn Asp Val Ser Ile His Glu Pro Ile Leu
AAACATAATTTATCCGAAATTCACCAG    AAT GAT GTC AGT ATT CAT GAA CCT ATA TTA

Thr Gly Leu Gly Asp Pro Ser Glu Tyr Lys Asp Lys
ACT GGT TTA GGA GAT CCT TCG GAA TAC AAA GAT AAG GTATTTTCCAACCTAATGATAAT

           Asn Val Leu Val Val Glu Asp Ile Ile Asp Thr Gly Lys Thr Ile
TTTTGTAAG  AAT GTT CTT GTG GTC GAA GAT ATA ATT GAC ACA GGA AAA ACA ATA

Thr Lys Leu Ile Ser His Leu Asp Ser Leu Ser Thr Lys Ser Val Lys Val Ala
ACG AAG CTC ATA AGC CAT TTG GAT AGT TTG TCT ACG AAA AGT GTT AAA GTC GCA

Se
AGGTATGTTTTGTAATATTTCCATCTTTCCAGTCTCATGATGTCAAAGTCGTCCCTCTATGTATCGTACTG

..approx. 1.0 kb...ATCAGAAGTAAGAATTACATGATAATTAGTATGGGAACAGAGAATAAGTTTC

ACTAACTTTGATAAAATTTGTAAAAATCAGTGTTTTTATCAAATCTTACATCATGAGGTTTCCGAAATCTG

             r Leu Leu Val Lys Arg Thr Ser Pro Arg Asn Asp Tyr Arg Pro
TTTTCTTT   AGC CTC CTC GTC AAG CGA ACA TCG CCC AGA AAT GAT TAC CGA CCA

Asp P
GAC TGTAAGTAAATTTTTATGACTTTTTAGGTTTGGTTTTTTGGTTTAGGCAAAATCATTCAATCTATGT

....approximately 2.8 kb.... CAACCAGTTCGGAGGACGAATATCCACTGGAATTATTCTTTG
```

```
CTCCATATTCGAGTAGATTTAGAAGTTTCTATAGGAAAGCATGAAATTCTATATATATTTTCTTATTTTTG

                             he  Val Gly Phe Glu Val Pro Asn Arg Phe
TTCCATTTCTTTTCCCGTTTGATCCTAAGTT  GTT GGT TTT GAA GTT CCA AAT CGA TTT

Val Val Gly Tyr Ala Leu Asp Tyr Asn Asp Asn Phe Arg Asp Leu His
GTC GTT GGC TAT GCT TTA GAC TAT AAT GAT AAT TTC CGT GAC CTG CAC GTGAGTT

ACCGTCTGCTACATCAACTTATAGAAGTTTGCTTATTTATTTTGCTGAATACTCTTGTGTGAAACTATGAA

.......................approximately 2.2 kb.....  CAACCATGGTTGTAGGTTTT

TTTATTATTTTAATGTTTTGTTTAGGAAAATTGTTGATTTACTTATGCTGAATTACCTCTTTTTTTCTTCT

                      His Ile Cys Val Ile Asn Glu Val Gly Gln Lys Lys
TTTTTTCTCACATTATTCAG  CAT ATT TGC GTG ATT AAT GAA GTG GGT CAA AAA AAA

Phe Ser Val Pro Cys Thr Ser Lys Pro Val OC
TTC TCT GTA CCC TGT ACA TCA AAA CCT GTT TAA ATGTTTTTTAGTTGAGATAATAACAAG

AAACTAAATAAATACAGAAACAGAATTTCCTTCAAAATGGTCATTCTGTAAATAATTAATCAATATCACAG

CCAAAAAAATATTTAACACTTCGTTGTTCTCTTCCAACTTCTTCTATACTTTTGTTATCCTATCTATACTA

TCTATACTACATATACTATGCTATCTATCCTATCTATATACTATACTATCGTTAGTCTATACTTGTATAAT

CAACATATCTTCAATAAGATATGTATTATTATCAACGAAACTGATTTTTCCATATCTATATATATTCTTAA

CTTATTTCTTTTTTTTATATTTTATTTCTAAGATCATATTTGGCACACAAATGTTTTTCTTTCTTTTTTCT

CTTTTTTTCTTGTTTTACTCCTTTCATTTCCGTTGATCAATTGATTAATCAATTTAATTGATTATTATTAC

ACTTCTTATACTTCCTTCCCCTCCCCCAACTTCTATCATCAAAATAAAATGTTTCTTCATAAGAAAGATTC

TAATGTTGTATGATTGACCAAAAAACAAAAAAAAACATTTTTCAATAATGATCATTAATTTCTAGTGTTAT

TTATTACTTATTAAAAATATTTTTTAGTATTATTGTTATTAATAAAACTAAAATCGAA↓TGTAAAATTGTT

TTCATTGTGTATGCGTTTTATAGTATTCTTTTGATAGTATTAAGGTTAGTTTATTGAAATGTAATAATAAG

CTTGCCCTGATACGGTTGAGAGTGGGAGAGTCAGCTCTCCTCTCGAATGCTCTCACATGGTCACGCGTAT
```

Figure 3. Partial sequence of the HGPRTase gene of *S. mansoni*. Within the partial sequence shown, 4 potential ATG start codons at the 5' end of the gene are underlined twice and consensus splice signals, are underlined once. An open reading frame begins well upstream of the third ATG codon, which has been proposed to represent the probable amino terminus of the schistosomal HGPRTase (4). A potential "TATA" sequence (21) is outlined at 102 bases upstream from the ATG codon that represents the probable amino terminus of the schistosomal HGPRTase. A complete "CAAT" box (46) is not found in the 5' leader sequence, although the outlined sequence AGCCAA that is 12 bp upstream of the potential "TATA" box has been identified as a "promotor binding site" (20). The arrow near the 3' end of the reported sequence shows where the cDNA sequence, which continues with a series of 15 A residues, diverges from the genomic sequence.

mansoni is composed of 8 exons and 7 introns distributed over approximately 8 kb of DNA. Analysis of the sequence upstream of the presumed start codon (4) for the schistosomal HGPRTase reveals the presence of a hexamer, AGCCAA, that has been shown by DNAase I footprint analysis to function as a "promoter binding site" for several eucaryotic genes (20). This hexamer is 114 bases upstream of the putative initiating methionine codon (Fig. 3). Also, a "TATA" consensus sequence, TATATAT (21), occurs only 12 bases downstream of the AGCCAA motif. The sequences

TABLE 1

			Codon	
Intron	Length	Splice Junction	Interruption	%A+G
1	31 bp	-GTA...CAG-	No	38
2	33 bp	-GTA...AAG-	Yes	54
3	42 bp	-GTA...CAG-	No	56
4	32 bp	-GTA...AAG-	No	43
5	~1.2 Kbp	-GTA...TAG-	Yes	?
6	~3.0 Kbp	-GTA...AAG-	Yes	?
7	~2.4. Kbp	-GTG...CAG-	No	?

TTTCAG and TTTTGTAG, which have been identified in nematode actin genes as 3' splice acceptor sequences for transplicing (22), are not found at the 5' end of the gene up to the first small intron. Pyrimidine rich sequences of 11 or more bases, typical of sequences found downstream of an adenine used for branch formation (6), are found just up-stream of the 3' splice sites of the sixth and seventh introns located nearest to the carboxyl terminus of the protein.

Intron structure

Our results show that introns 1-4 of the HGPRTase gene of *S. mansoni* are each only 31, 33, 42, and 32 bp in length, respectively (Table 1). In contrast to the clustering of the first 5 exons, the four exons closest to the 3' end of the gene (introns 4-7) are separated by introns each larger than 1.2 kb (Table 1). Canonical sequences for splice site recognition are found, as all of the introns begin with GTR (R = A or G) and end with AG. Three of the introns (#'s 2, 5 & 6) interrupt the codons of flanking exons, whereas the remaining four do not (Table 1). Preliminary data from RNA mapping experiments (results not shown) further confirms the existence of the 31 and 33 bp introns (introns 1 & 2).

DISCUSSION

The 3 introns of 31, 32, and 33 bp are the smallest ever described in a non-viral nuclear gene coding for protein. Although a 31 base intron has been reported to be removed from a late lytic RNA of simian virus 40 (23), Fu *et al.* (24) discuss size limitations for SV 40 introns in the small-t antigen pre-mRNA that are consistent with those reported for eucaryotic nuclear genes coding for protein (6,15,16). Furthermore, in contrast to the current dogma that a pyrimidine rich sequence of 11 or more bases is required for splicing (6,9), none of the 4 very small introns (vsi's) possess a pyrimidine sequence longer than 7 bases and the second and third vsi's are com-posed of more than 50% purine bases distributed throughout their length (Table 1). Also, the 4 vsi's lack the TACTAAC sequence used for branch formation in yeast and only the first and third introns have the YNYTRAY (R = purine base, Y = pyrimidine base, and N = A, T, G, or C) sequence used for branch formation in vertebrates (7,9,14). Computer analysis indicates the absence of significant secondary structures within single stranded transcripts of any of the 4 vsi's, in sharp contrast to secondary structural characteristics of self-splicing introns (6,14,25-27). These results suggest that other, as yet-undiscovered mechanisms may exist for the splicing of nuclear mRNA precursors in *S. mansoni.*

Although splicing is typically viewed as the joining of two exons from a single co-valent precursor, the trans-splicing of exons from two separate precursor molecules might be important in certain biological systems (6,7). For example, the parasitic protozoan, *Trypanosoma brucei,* appears to utilize trans-splicing as a natural step in the processing of mRNA precursors (28-30). In addition, the self-splicing of RNA precursors has been described in mitochondrial RNAs, in the pre-rRNA of *Tetrahymena pyriformis* (6,27) and in virusoid RNA's (31). Although certain identified sequences and secondary structures may play a role in self splicing (26), it is not known whether there is a minimal length requirement for self splicing introns.

Are the vsi's of *S. mansoni*, a primitive metazoan flat worm, unique or are they also characteristic of other metazoan genes? The free living round worm, *Caenorhabditis elegans*, is notable for having a number of genes with relatively small introns. For example, the collagen genes possess introns of 47 and 52 bp (32), the vitellogenin gene has 4 introns ranging from 47 to 70 bp (33), the glyceraldehyde-3-phosphate dehydrogenase gene has introns of 47 and 51 bp (34), and the heat shock genes have introns of 52 and 55 bp (35). Furthermore, introns 38-52 bp long are present within the sequence reported for the myosin heavy chain gene of *C. elegans* (36). One possible way for *S. mansoni* and *C. elegans* to splice small introns would be to have smaller SnRNPs. However, Thomas *et al.* (37) report that the SnRNPs of *C. elegans* are comparable to other SnRNPs.

With respect to intron positions within the coding sequence for HGPRTase, the fact that 3 of the 7 schistosomal introns interrupt flanking codons is consistent with the previous observations that almost half of the eucaryotic introns interrupt neighboring codons (38). The positions of 6 of the 7 introns within the coding sequence for *S. mansoni* HGPRTase are identical to the positions for 6 of the introns of mammalian genes (39). With respect to the one exception, the second intron occurs at the same position in the aligned sequences but there is a 3 amino acid addition to the schistosomal sequence. All but two of the mouse introns are larger than 2.9 kb (39). The two exceptions, near the carboxyl terminus of the protein, are approximately 200 and 800 bp long and correspond to schistosomal introns that are roughly 3.0 and 2.4 kb in length, respectively. An eighth intron in the mouse gene, which splits the putative phosphoribosylpyrophosphate (PRPP) binding domain between the fifth and sixth exons of mammalian genes, is entirely absent from the schistosomal gene. The fact that 7 of the 8 mammalian introns occur at identical locations within the coding sequence for the schistosomal HGPRTase indicates that the introns were probably present in the common ancestor to vertebrates and trematodes.

Some introns may be extraordinarily ancient (40). However, the appearance, subsequent elimination, or changes in size and sequence of introns may be an ongoing process in evolution. We know that introns can be excised, at least in mammals, and that they can probably be inserted, at least in the genes of invertebrates (41). With respect to evolutionary significance, Lonberg and Gilbert (42) have suggested that an intron within genes coding for a family of mononucleotide binding proteins may have been involved in bringing together the ancestors to building blocks (segments of the protein) that are found within contemporary members of this family of genes. These building blocks contain a β–α–β secondary structure which is believed to form the mononucleotide binding fold (mnbf). The mnbf is a common feature of a number of proteins, including the HGPRTase of *S. mansoni* (4). The second intron of the *S. mansoni* gene is located at nearly the identical position to that identified by Lonberg and Gilbert (42), within a predicted alpha helical structure of the putative mononucleotide binding fold. The coincidence of position of this intron is supportive of Lonberg and Gilbert's suggestion that an intron may have been involved in the evolutionary assembly of the mnbf.

Craik et al. (43,44) have suggested that intron-exon junctions tend to map at protein surfaces and that movement of splice sites would introduce changes in amino acid sequence at the surface of proteins. In the HGPRTase gene of *S. mansoni*, the location of the 3' splice site of the second intron differs by 9 bp from that of the mouse gene and results in the addition of 3 amino acids to the schistosomal protein. Hydrophobicity analysis of the amino acid sequence for the HGPRTase of *S. mansoni* (4) reveals that 6 of the 7 introns, including the second, are located within hydrophilic domains or at the interface between hydrophilic and hydrophobic domains of the protein. This physical data is consistent with the hypothesis that intron position may have provided a means to introduce variability at the surface of proteins (44).

In light of these concepts that introns may provide a mechanism for introducing variability into proteins (44), one cannot help but wonder what is the evolutionary significance of the organization of the single copy schistosomal HGPRTase gene. The gene for the HGPRTase of *Plasmodium falciparum* has been isolated and partially sequenced (45). The coding sequence indicates no intron interrupts the amino acid sequence which aligns with positions 30 to 218 of the mouse enzyme (45). Since intron positions sometimes predate the divergence of plants and animals (10,11), the absence of introns from the malarial HGPRTase gene might represent a loss of introns. Like schistosomes, malarial parasites also lack *de novo* purine nucleotide biosynthesis and are thus dependent on HGPRTase. In the schistosome HGPRTase gene, one intron is missing and the remaining four very small introns are flanked by exons encoding the putative critical protein domains: the mononucleotide binding fold, the purine binding site and the PRPP binding site (Fig. 1). If introns allow increased likelihood of exon rearrangements (42) or alterations in protein surface properties (43,44) then reduction in intron size or the elimination of introns in the HGPRTase gene might have evolved to reduce the chances for mutational change in this enzyme which is so critical to survival of the parasite.

Splicing mechanisms known to operate in mammalian genes (6), that are believed to require a minimal intron size of approximately 50 bp are unlikely to apply to the schistosomal HGPRTase gene and possibly to other metazoan genes as well. Other splicing mechanisms must now be sought. It has been suggested that the earliest introns may have been self-splicing (25). Sharp has speculated that a step in the transition from the self-splicing intron to the nuclear intron would be the appearance of transacting factors (i.e. catalytic RNA) that execute the excision of the introns, thereby freeing the sequences within a typical intron from the constraints associated with self-splicing. The trans self-cleavage of viroid RNA (31) seems to fit this model (25) and may partially fulfill the role as a missing link in intron evolution. In the absence of known mechanisms for splicing the vsi's from transcripts of the schistosomal HGPRTase gene, it is tempting to suggest that the vsi's might also be removed from transcripts of this gene by a "trans self-cleavage" mechanism. Whether or not the schistosomal vsi's are in fact a "missing link" in intron evolution, the intron/exon structure of the schistosomal gene, compared to the genes for the human and malarial enzymes, shows that there are important evolutionary changes in intron size among primates, primitive metazoa, and protozoa.

ACKNOWLEDGEMENTS

We thank G. R. Newport for providing bacteriophage libraries of genomic DNA from *S. mansoni*, K. C. Lim for maintaining life cycle stages of *S. mansoni*, K. Shen for performing some of the DNA sequencing, A. Eakin for careful reading of the manuscript, and C. Craik for helpful discussions and editorial suggestions. Supported by NIH Grant Number AI-24011. C. C. Wang is a Burroughs Wellcome Scholar in Molecular Parasitology.

*EMBL accession nos X13531 – X13534 (incl.)

REFERENCES

1. Kelley, W. N., Greene, M. L., Rosenbloom, F. M., Henderson, J. F. and Seegmiller, J. E. (1969) *Ann. Intern. Med.* **70**, 155-206.
2. Seegmiller, J. E., Rosenbloom, F. M. and Kelley, W. N. (1967) *Science* **155**, 1682-1684.
3. Davidson, R. L. ed., *Somatic Cell Hybridization*, Raven Press, New York (1974).
4. Craig, S. P. III, McKerrow, J. H., Newport, G. R. and Wang, C. C. (1988) *Nuc. Acids Res.* **16**, 7087-7101.
5. Mott, K. E. (1984) *Schistosomiasis: New goals.* World Health, World Health Organization, Geneva.
6. Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S. and Sharp, P. A. (1986) *Ann. Rev. Biochem.* **55**, 1119-1150.
7. Sharp, P. A., Konarska, M. M., Grabowski, P. J., Lamond, A. I., Marciniak, R. and Seiler, S. R. (1987) *Cold Spr. Harb. Symp. Quant. Biol.* **52**, 277-285.
8. Breathnach, R. and Chambon, P. (1981) *Ann. Rev. Biochem.* **50**, 349-383.
9. Ohshima, Y. and Gotoh, Y. (1987) *J. Mol. Biol.* **195**, 247-259.
10. Zakut, R., Shani, M., Givol, D., Neuman, S., Yaffe, D. and Nudel, U. (1982) *Nature (London)* **298**, 857-859.
11. Gilbert, W. (1985) *Science* **228**, 823-824.
12. Jhiang, S. M., Garey, J. R. and Riggs, A. F. (1988) *Science* **240**, 334-336.
13. Rio, D. C. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2904-2908.
14. Krainer, A. R. and Maniatis, T. (1988) *RNA splicing,* in *Transcription and Splicing,* Hames, B. D. and Glover, D. M. eds., IRL Press, Oxford, U.K. and Washington, D.C., 131-206.
15. Wieringa, B., Hofer, E. and Weissmann, C. (1984) *Cell* **37**, 915-925.
16. Thompson-Jager, S. and Domdey, H. (1987) *Mol. Cell Biol.* **7**, 4010-4016.
17. Rautmann, G., Matthes, H. W. D., Gait, M. J. and Breathnach, R. (1984) *EMBO J.* **3**, 2021-2028.
18. Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982) *Molecular Cloning, A Laboratory Manual*, Cold Spring Harbor Laboratory Press, NY.
19. Martinez, H. M. (1988) *Nuc. Acids. Res.* **16**, 1789-1798.
20. Jones, K. A., Kadonaga, J. T., Rosenfeld, P. J., Kelly, T. J. and Tjian, R. (1987) *Cell* **48**, 79-89.
21. Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C. and Chambon, P. (1980) *Science* **209**, 1406-1414.
22. Krause, M. and Hirsh, D. (1987) *Cell* **49**, 753-761.
23. Ghosh, P. K., Reddy, B. B., Swinscoe, J., Lebowitz, P. and Weissman, S. M. (1978) *J. Mol. Biol.* **126**, 813-846.
24. Fu, X.-Y., Colgan, J. D. and Manley, J. L. (1988) *Mol. Cell. Biol.* **8**, 3582-3590.
25. Sharp, P. A. (1985) *Cell* **42**, 397-400.
26. Sharp, P. A. (1987) *Cell* **50**, 147-148.
27. Cech, T. R. (1987) *Cell* **44**, 207-210.
28. Boothroyd, J. C. and Cross, G. A. M. (1982) *Gene* **20**, 281-289.
29. Parsons, M., Nelson, R. G., Watkins, K. P. and Agabian, N. (1984) *Cell* **38**, 309-316.
30. Guyaux, M., Cornelissen, A. W. C. A., Pays, E., Steinert, M. and Borst, P. (1985) *EMBO J.* **4**, 995-998.
31. Forster, A. C., Jeffries, A. C., Sheldon, C. C. and Symons, R. H. (1987) *Cold Spr. Harb. Symp. Quant. Biol.* **52**, 249-260.
32. Kramer, J. M., Cox, G. N. and Hirsh, D. (1982) *Cell* **30**, 599-606.
33. Spieth, J., Denison, K., Zucher, E. and Blumenthal, T. (1982) *Nuc. Acids Res.* **13**, 7129-7138.

34. Yarbrough, P. O., Hayden, M. A., Dunn, L. A., Vermersch, P. S., Klass, M. R. and Hecht, R. M. (1987) *Biochim. Biophys. Acta* **908**, 21-33.
35. Russnak, R. H. and Candido, E. P. M. (1985) *Mol. Cell. Biol.* **5**, 1268-1278.
36. Karn, J., Brenner, S. and Barnett, L. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 4253-4257.
37. Thomas, J. D., Conrad, R. C. and Blumenthal, T. (1988) *Cell* **54**, 533-539.
38. Traut, T. W. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2944-2948.
39. Melton, D. W., Konecki, D. S., Brennand, J. and Caskey, C. T. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 2147-2151.
40. Quigley, R., Martin, W. F. and Cerff, R. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2672-2676.
41. Rogers, J. (1985) *Nature* **315**, 458-459.
42. Lonberg, N. and Gilbert, W. (1985) *Celll* **40**, 81-90.
43. Craik, C. S., Sprang, S., Fletterick, R. and Rutter, W. J. (1982) *Nature* **299**, 180-182.
44. Craik, C. S., Rutter, W. J. and Fletterick, R. (1983) *Science* **220**, 1125-1129.
45. Sullivan, M. A., Lloyd, D. B. and Holland, L. E. (1987) in *Molec. Strategies of Parasitic Invasion*, Agabian, N., Goodman, H. and Nogueira, N. eds., Alan R. Liss, Inc. N.Y., 575-584.
46. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) *Nuc. Acids Res.* **8**, 127-142.