RESEARCH NOTE

Characterization of an Abundant *Schistosoma mansoni* Transcript with No Homologs in the Databases

Wendell SF Meira, Glória R Franco, Élida ML Rabelo*, Sérgio DJ Pena/+

Departamento de Bioquímica e Imunologia
*Departamento de Parasitologia, ICB-UFMG, Av. Antônio Carlos 6627, 31270-010 Belo Horizonte, MG, Brasil

Key words: *Schistosoma mansoni* - expressed sequence tags - database searches - reverse transcribed-polymerase chain reaction

The expressed sequence tag (EST) approach that we have used in the *Schistosoma mansoni* Genome Project is a powerful technique for the discovery of new genes of the parasite (GR Franco et al. 1995 *Gene 152*: 141-147, E Dias Neto et al. 1997 *Gene 186*: 135-142). In a recent comparative study of gene expression in distinct developmental stages of the parasite life cycle using the EST strategy, we identified 466 different genes. From this total, 427 were novel and 333 of them could not be identified based on homologies with database sequences (GR Franco et al. 1997 *DNA Research 4*: 231-240). The high frequency of some of these "unknown" genes in different cDNA libraries suggests that they might have important roles in the biology of *S. mansoni* and thus may constitute possible targets for drug design or vaccine production. One of these genes, highly abundant in one of four adult worm libraries that we are studying in our laboratory, was selected for further characterization.

This research was funded by grants from UNDP/WORLD BANK/WHO Special Programme for Research in Tropical Diseases, Fapemig and CNPq/PADCT.
+Corresponding author. Fax: +55-31-441.5963. E-mail: spena@dcc.ufmg.br
Received 4 May 1998
Accepted 31 August 1998

After clustering analysis of ESTs from different *S. mansoni* cDNA libraries using the program ICATOOLS (Franco et al. 1997 *loc. cit.*), we identified a cluster composed of 16 ESTs from both cDNA ends that corresponded to an unknown gene highly frequent in an adult worm cDNA library. We have called this gene *AUT1* for abundant unknown transcript 1. A single strand consensus of approximately 1.6 kb long was derived from the alignment of the 16 EST sequences using the program DNAsis. In order to obtain the cDNA full-length sequence from both strands, the cDNA clone containing the largest insert was digested with the restriction enzymes *Hin*dIII and *Sph*I (Fig. 1), the fragments produced were further cloned into the pUC18 vector (Pharmacia) and completely sequenced on both strands using the Thermo-Sequenase Fluorescent Labeled Primer Cycle Sequencing kit (Amersham Life Science) and the A.L.F. Automated DNA Sequencer (Pharmacia). Another strategy used to obtain the cDNA sequence from both strands was the amplification by polymerase chain reaction (PCR) of different regions of the cDNA using specific primers designed for the gene. The fragments were cloned into the pUC18 vector (*Sma*I cloning site) using the SureClone Ligation kit (Pharmacia) and sequenced as before. The sequences generated from both processes were aligned using the DNAsis program and the cDNA full-length sequence from both directions was obtained, totaling 1520 bp (Fig. 2). The cDNA was translated into the six possible frames and the length of the longest open reading frame (ORF) was 1005 bp long, potentially encoding a protein of 335 amino acids (Fig. 2).

Analysis of the primary structure of the putative protein coded by *AUT1* gene shows one potential site for N-linked glycosylation, as well as eight sites for phosphorylation by protein kinase C, five for phosphorylation by casein kinase II and eight sites for cAMP-dependent protein kinase, suggesting this protein might be phosphorylated in the organism. The protein does not contain any signal sequence responsible for translocation across the endoplasmic reticulum (ER) membrane, as seen in secretory and plasma membrane spanning proteins, or stretches of hydrophobic residues for plasma membrane insertion. Several searches were performed on distinct databases of protein sequences, typical protein domains and families of proteins as an attempt to identify the predicted protein or a specific domain on it. All the searches provided neither homology with any sequences in databases nor identifiable structural domains.

Primers designed for PCR amplification of a fragment containing the complete coding region

of the cDNA were used for amplification of the *AUT1* gene from the parasite genome. The 1.75 kb fragment obtained was cloned into pUC18 vector. Segments of this insert were amplified using the same primers described before for the amplification of the cDNA. These segments were also subcloned into pUC18 and sequenced in both directions. The generated sequences were aligned using the program DNAsis, given a 1754-bp genomic sequence. The gene contains three introns. The first intron (129 bp) interrupts the coding region of the gene and the other two introns (42 bp and 63 bp) are located towards the 3' untranslated region (Fig. 1). The canonical donor/acceptor splice sites are conserved at all exon/intron junctions.

The expression of the *AUT1* gene in different stages of the parasite life cycle was verified by reverse transcribed PCR (RT-PCR). A 498-bp fragment containing the total 3' untranslated region and the final part of the coding region was amplified from mRNA preparation from eggs, miracidia, 3h-schistosomula and adult worms. The gene was seen to be expressed in all stages analyzed (data not shown).

In summary, we have characterized an unknown gene of *S. mansoni*, (*AUT1*) that is expressed in different stages of its life cycle. Although the putative protein coded by this gene may potentially be phosphorylated, it does not possess any characteristic structural domain, membrane spanning regions, signals for localization in the nucleus or organelles or signal for translocation across the ER membrane. Homology searches conducted in distinct databases did not show any significant similarity to existing genes, proteins or ESTs. This investigation is being continued in our laboratory. The protein is already expressed in bacteria and polyclonal antibodies are going to be produced for immunolocalization assays, so that we can have some clue of the cellular function of this protein.

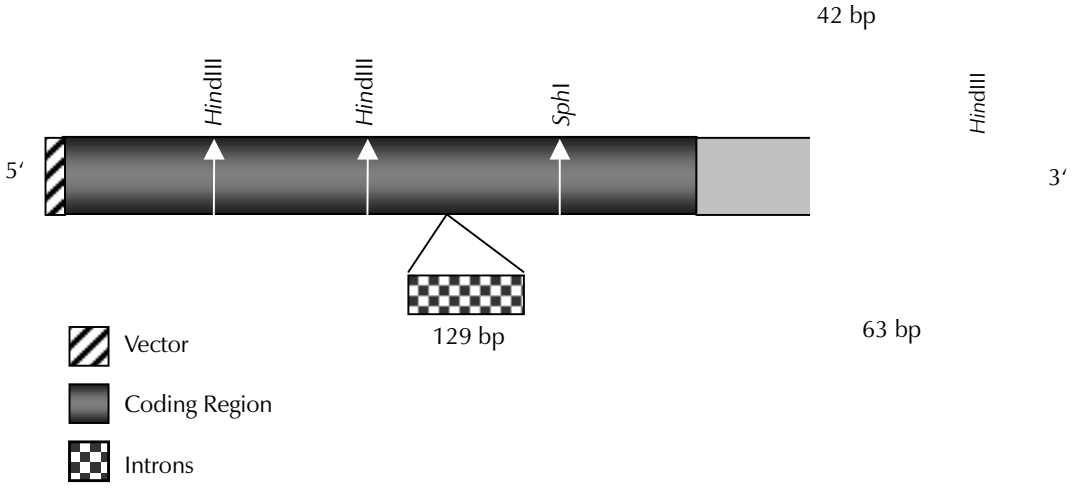

Fig 1: schematic representation of the *AUT1* gene. Positions and sizes of the three introns, the coding region and the restriction sites used for digestion and subcloning are indicated.

```
1    GGT GAT CAT GTT GAC TGC TTG TCC ACA GTG TGT TCT GTA GTC ATG GCT GGT GTT   54
1                                                        M   A   G   V     4

55   GAC CAC GCC TAT CAG TAC TCC GTG GAA GTG AAG AGC AGG TTC TCA CTG TTC CTG  108
5     D   H   A   Y   Q   Y   S   V   E   V   K   S   R   F   S   L   F   L   22

109  GAC GAC ACT CTG AAT AGT GAA GAC CCA GAT ATC CTG CTA TCC AAG TTG CAG AGT  162
23    D   D   T   L   N   S   E   D   P   D   I   L   L   S   K   L   Q   S   40

163  AAA CGA GGT GAA AAA ACT AAA AAG GAT AAG CCT CAT TTG CAA CAG CAA CAT GTG  216
41    K   R   G   E   K   T   K   K   D   K   P   H   L   Q   Q   Q   H   V   58

217  GCA CCT ACA AAG GCC GAC ACG ATA ACT AAG AGT GAG GTG AAG CTT GAT ACT GCA  270
59    A   P   T   K   A   D   T   I   T   K   S   E   V   K   L   D   T   A   76

271  ACT CCG AAA GCT GGT TCT CGG GTG TCG AAA ACC CCC AAC TCC ACA GAG CCG CCA  324
77    T   P   K   A   G   S   R   V   S   K   T   P   N   S   T   E   P   P   94

325  CCA GTG CCA CCT GAA GAC GTG CAA ATT ACA TCC GCT AAA GGG ACA GAT GAG CCA  378
95    P   V   P   P   E   D   V   Q   I   T   S   A   K   G   T   D   E   P  112

379  ATT AGT ACC TTT NTN CGT GGT CGT GGC TCA GGT AGG GGA ACA CCT CGA GGA ATG  432
113   I   S   T   F   X   R   G   R   G   S   G   R   G   T   P   R   G   M  130

433  CGT GTA GGT AGA GGG CAG GGG CCA CGA ATA GCT CCG ACA GAA GCA CCT CAA GAT  486
131   R   V   G   R   G   Q   G   P   R   I   A   P   T   E   A   P   Q   D  148

487  TCA GTG AGT GAT TTG AAT GCT CCC AGA GGG TCA AGC TTT GAG CCT AGA GGA CGT  540
149   S   V   S   D   L   N   A   P   R   G   S   S   F   E   P   R   G   R  166

541  GGG AGA GGC AGA GGA AGA GGA ATG TTT GGT CGA GGT CGT GGG ATG CCA TTC AAT  594
167   G   R   G   R   G   R   G   M   F   G   R   G   R   G   M   P   F   N  185

595  TCA AAT CGA GAC TTC GAA AAT CAG GAT GGC CCA GAT CGT CAG GGT CCA AGA CAA  648
185   S   N   R   D   F   E   N   Q   D   G   P   D   R   Q   G   P   R   Q  202

649  TAT GGT CGA AGA GAT GGG AAT TGG AAT TCT CAG GAT GTT GAT GGT CTA ATA ATG  702
203   Y   G   R   R   D   G   N   W   N   S   Q   D   V   D   G   L   I   M  220

703  CCT GAG AGT GGT GAC TCA GAG CAA GTT GTG CGT TTC GCA GAC GAC CGT AAC GAA  756
221   P   E   S   G   D   S   E   Q   V   V   R   F   A   D   D   R   N   E  238

757  GTC GAG GAC CAG CCA GAG CAT GCC ACA GCG GAG AAT GAG GAG GGT GTG GTT GTC  810
239   V   E   D   Q   P   Q   H   A   T   A   E   N   E   E   G   V   V   V  256

811  GGC ACA GAA ACC CCA GTT GAG GAA GAA CCG AAA AGC TAC ACA CTA GAA GGA TAC  864
257   G   T   E   T   P   V   E   E   E   P   K   S   Y   T   L   E   G   Y  274

865  AAA GCT ATG CGT CAG TCT TCC AAA CCA GCC GTT TTA TTA AAC AAC AAA GGA CTA  918
275   K   A   M   R   Q   S   S   K   P   A   V   L   L   N   N   K   G   L  292

919  CGC AAA GCT AAT GAC GGC AAA GAT GTG TTT GCA AAC ATG GTG GCT CAC AGA AAA  972
293   R   K   A   N   D   G   K   D   V   F   A   N   M   V   A   H   R   K  310

973  TTA CAA GAG GTC TCC GAA GAT GTC TAC GAG GTG GAG GAG AGG AAA ACG AGC CTG 1026
311   L   Q   E   V   S   E   D   V   Y   E   V   E   E   R   K   T   S   L  328

1027 AGA GCA TCA GTC GAT CGA TAT TGA TTT CCC TTT CGC TGA CGA CTT TGG AAA CCG 1080
329   R   A   S   V   D   R   Y  ***  ...  ...  ...  ...  ***  ...  ...  ...  ...  ...  335

1081 GTG CAG GGC CAA GAA GTG AAC GTG GGG GCG AGG AAG AGG TCA GAT AAG GAG TGA 1134
1135 CGG AAG AGG GCC GTG GTT TCG GCC GTG GTT TGC CTA TTC CAA CGA GTC ACA CTG 1188
1189 CGA CCT CCA GCA ATT AAC AAC GAC GCC AGG AGT TCC CTT CCT TGA AGT GAA ATC 1242
1243 TCA AAG GAA AAT GAA AAC AGA TTG TTG GGT TGA TTA TAT TCT TCC TTG TCA ACT 1296
1297 TTT ACC TTT CTC TAA GGC TTT TAG TCC AAG CCT CTA CTT ACT CAT ACT CAT CTT 1350
1351 TCC CTA TTA AAC TTC AGC TTT GTA ATC TAG TTT CTC CTA AAA GCA GTT CAA GGA 1404
1405 CTG GAT TTG GCT GTG CTG CTG GAG ATC TAA TTG CAA AGC TTT GTA CAT GAG TGT 1458
1459 TTT CAG AAT GTC AGT TTT GCG GTT GAA GTG ACC AAT CGC CCC AAA TAA AAA TTA 1512
1513 ATA AGA TG 3'                                                        1520
```

Fig 2: nucleotide sequence of the cDNA and the putative protein encoded by the *AUT1* gene. The potential sites for protein phosphorylation by PKC - {ST}X{RK}, CKII - {ST}XX{DE} and cAMP-dependent kinase - {RK}XX{ST} are underlined in the sequence. The potential site for N-glycosylation (NX{ST}) is in bold italics.