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.

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 *Hind*III 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-Sequnase 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

---

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
of the cDNA were used for amplification of the \textit{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 \textit{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 \textit{S. mansoni}, \textit{(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 \textit{AUT1} gene. Positions and sizes of the three introns, the coding region and the restriction sites used for digestion and subcloning are indicated.
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.
Characterization of Unknown *S. mansoni* Gene • Wendel SF Meira et al.