Volume 15 Number 14 1987 Nucleic Acids Research

Sequences of two genomic fragments containing an identical coding region for a putative egg-shell precursor protein of *Schistosoma mansoni*

Werner Kunz, Karin Opatz, Marion Finken and Peter Symmons

Institut für Genetik der Universität Düsseldorf, Universitätsstrasse 1, D-4000 Düsseldorf 1, FRG

Submitted June 16, 1987 Accession no. Y00374

We have sequenced two genomic DNA fragments corresponding to a putative eggshell protein precursor of the human parasite S. mansoni (1). The 1144 bp HindIII-EcoRI and 901 bp EcoRI-EcoRI fragments were isolated from a Liberian strain of S. mansoni. The 635 bp 3' to the translation start point of both fragments are perfectly homologous. 220 bp upstream from the start of translation are also highly conserved (identity indicated by ·, differences by *). Further upstream divergence increases although homology is still discernible. Two overlapping TATA boxes (outlined) place the likely start of transcription (») 40 bp upstream from the start of translation and indicate that there are no introns (compare cDNA (1)).

```
     HindIII-EcoRI fragment
   1 AAGCTTGTAAATAAAATCGGTGCATTCTCCTCATGATACACTGGTCAACACTCAGTCTTTAATATCAATTTAAGTGTGTACAACTGATGAGTCCCATTCAAAATGA
 107 AACACTTATCCTAGATTCATATGCAGCTGCTGTAAAAAAACCTTCAAATACATGTTAGTGAAGTTATAAGTTAGCATAAAATTTAGATTAATGATAAGTGGATCAA
 213 ATAAATGCGTTGAACTTATGTAGATACAGGAAGAAAGACTGAAGTTTCATCAAGGACATTGATTGTAAAATGTATGTACAGTTTCATTGATATTTTTAACTATCAC
                                        ********  ***     ***** *** *******  *  **  *          *       *
     EcoRI-EcoRI fragment            ···TTCTAACG··AGA·····AATAG·GAG·AAATTCG··A··TA··T··········C·······a·······

 319 GCTCAACTATCATTAACG-TTACACAAAAGACAGTTCGACAGTAGAGTGCACAtGCGTAATATGTATTGAATAATGATGCACTTAGTGAGGCACAACTCTTCAAAT
                       *                                  * *
     ··················A··································T·c··················································
                                                                                             1
                                                    »                                        M   K   Q   S   L
 425 CTATAATCAAAAACAATATATAAATGATAAATCACACTAGTCTACACATCATCACACCCAGTACAACAACACCAACAATTTGAAAA ATG AAA CAG TCA CTC
                                                                                          *
     ·····································································································C(identical till end)
                      10                                      20                                      30
      T   L   V   F   L   V   A   I   G   Y   A   T   A   H   T   T   S   H   D   Y   S   G   G   Y   G   G
 526 ACA CTC GTC TTC TTA GTA GCC ATT GGT TAC GCC ACC GCC C¹AC ACC ACA TCA CAT GAC TAT TCG GGT GGG TAC GGT GGC
                                      40                                      50
      G   C   Y   G   S   D   C   D   S   G   Y   G   D   S   G   Y   G   G   G   C   T   G   G   D   C   G
 604 GGT TGC TAT GGT AGC GAT TGT GAT AGC GGT TAT GGC GAT AGT GGA TAT GGT GGA GGC TGT ACT GGT²G³GT GAC TGT GGC
              60                                      70                                      80
      G   G   Y   G   G   G   C   S   G   G   D   C   G   N   Y   G   G   G   Y   G   G   D   C   N   G   G
 682 GGC GGC TAT GGT⁴GGA GGT TGC AGT GGT GGA GAT TGT GGT AAT TAC GGT GGT GGC TAT GGT GGT GAT TGC AAT GGT GGA
                              90                                     100
      D   C   G   N   Y   G   G   G   Y   G   G   G   N   G   G   G   C   S   G   G   N   C   G   G   G   F
 760 GAT TGT GGT AAT TAC GGT GGT GGC TAT GGT GGT GGG AAT GGT GGT GGT TGC AGT GGT GGC AAT TGT GGA GGT GGC TTC
     110                                     120                                     130
      D   E   A   F   P   A   P   Y   G   G   D   Y   G   N   G   G   N   G   F   G   K   G   G   S   K   G
 838 GAT GAG GCC TTC⁵CCT GCC CCC TAT GGC GGT GAT TAT GGT AAC GGT GGC AAC GGC TTT GGA AAA GGT GGT AGT AAA GGC
                     140                                     150                                     160
      N   N   Y   G   K   G   Y   G   G   G   S   G   K   G   K   G   G   G   K   G   G   K   G   G   K   G
 916 AAC AAT TAT GGA⁶AAG GGT⁶TAT GGC GGT GGT AGC GGT⁷AAG GGT⁷AAG GGT GGT GGC AAA GGT GGC AAA GGC GGC AAA GGT
                                     170
      G   T   Y   K   P   S   H   Y   G   G   G   Y   -
 994 GGC ACT TAC AAA CCC AGC CAT TAT GGA GGC GGT TAC TGA GGCACCAGTTGAGTTGTGGATCATTCTAATTTGTTTGTGTCACACTCTCCA
1084 CTGTCCTATTTTTCTACACACCTCTCAATTCAACTCACTGTAATATAGTCGTGTTTGAATTC
```

Lower case characters indicate ambiguous bases in sequence autoradiograms. Published sequence differences: in (1) ^{1}=T, ^{2}=C, ^{3}=A, ^{5}=A (point substitutions), ^{4}=GGTGGCTATGGT (insertion); in (2) 24 bp are missing, whose exact position is ambiguous because of repeats, ^{6} and ^{7} indicate the limits of the 5'and 3'borders respectively.

REFERENCES

1. Bobek, L., et al. (1986) PNAS 83, 5544-5548.
2. Simpson, A.J.G., et al. (1987) Mol. Biochem. Parasitol. 22, 169-176.