

# Complete sequence of the gene encoding the largest subunit of RNA polymerase I of *Trypanosoma brucei*

Waldemar Jess, Andrea Hammer and Albert W.C.A. Cornelissen

*Max-Planck-Institut für Biologie, Molecular Parasitology Unit, Spemannstrasse 34, 7400 Tübingen, FRG*

Received 20 March 1989

We have set out to clone the trypanosomal gene encoding the largest subunit of RNA polymerase I. We screened a genomic library with a synthetic oligonucleotide probe encoding an eleven amino acid sequence motif, YNADFDGDEM<sub>N</sub>, which has been found in all eukaryotic RNA polymerase largest subunit genes analyzed so far. We isolated the Trp11 locus and determined the complete sequence of the gene encoded within this locus. The deduced amino acid sequence contains the highly conserved RNA polymerase domains as well as the previously identified RNA polymerase I-specific hydrophilic insertions. Therefore, the gene most closely resembles the largest subunit of RNA polymerase I.

RNA polymerase I; Largest subunit; Nucleotide sequence; Sequence homology; (*Trypanosoma brucei*)

## 1. INTRODUCTION

Eukaryotic RNA polymerases are multi-subunit enzymes responsible for transcription. Three classes of RNA polymerases have been described, which are named I(or A), II(or B) and III(or C). The nucleolar enzyme RNA polymerase I transcribes the ribosomal RNA genes, II is responsible for the transcription of all mRNA precursors and III transcribes the small RNA genes, encoding e.g. 5 S and tRNAs (reviewed in [1,2]).

Protein-coding genes are transcribed by RNA polymerase II which is, in most eukaryotes, characterized by its high sensitivity to the toxin  $\alpha$ -amanitin [3]. This contrasts strongly with the transcription of the VSG and the physically linked ESAG genes of the African trypanosome, *Trypanosoma brucei* [4–7]. Transcription of these two gene classes is completely insensitive to  $\alpha$ -amanitin [5]. Other trypanosomal protein-coding genes are, however,

transcribed by an RNA polymerase II which is sensitive to  $\alpha$ -amanitin [8,9].

To explain the unusual transcription of the VSG and ESAG genes, two alternative hypotheses have been proposed. Firstly, the eukaryotic RNA polymerase resistant to  $\alpha$ -amanitin, RNA polymerase I, might transcribe these genes despite their protein-coding nature. This hypothesis is supported by the observation that the putative promoter region of two VSG transcription units shows homology with the consensus sequence of the eukaryotic RNA polymerase I promoter [10,11]. Secondly, the VSG transcription unit might be transcribed by a modified RNA polymerase II, which has gained resistance towards the toxin  $\alpha$ -amanitin. The latter hypothesis is supported by the fact that *T.brucei* contains two slightly different copies of the Pol II gene [12]. In all other eukaryotic species analyzed so far, Pol II is encoded by a unique gene. Moreover, genetic analysis in higher eukaryotes has demonstrated that mutations conferring  $\alpha$ -amanitin resistance to RNA polymerase II have all occurred in the largest subunit of RNA polymerase II [1,2]. Therefore, the presence of the second copy in trypanosomes might generate a modified RNA polymerase II, transcribing VSG and ESAG genes.

To obtain additional information on the RNA

*Correspondence address:* A.W.C.A. Cornelissen, Max-Planck-Institut für Biologie, Molecular Parasitology Unit, Spemannstrasse 34, 7400 Tübingen, FRG

*Abbreviations:* VSG, variant surface glycoprotein; ESAG, expression site-associated gene; Pol I, Pol II and Pol III indicate the largest subunit of the respective RNA polymerase

polymerase transcribing the VSG transcription unit, we wanted to extend our analysis of trypanosomal RNA polymerases [12,16] to Pol I. We have previously reported the isolation of the locus, Trp11, most likely encoding Pol I [12]. In this study we report the isolation and the complete sequence of the gene encoding Pol I. Our aim is to use this gene as a tool to examine the transcription of VSG genes in further detail, allowing us to critically evaluate the two alternative hypotheses described above.

## 2. MATERIALS AND METHODS

### 2.1. Constructing and screening of the EMBL3 library

The genomic library of *T.brucei* in EMBL3 was prepared and screened with a synthetic oligonucleotide probe <sup>5</sup>TAC/T AAT GCT GAC/T TTC GAC/T GGT GAC GAA ATG AAT<sup>3</sup> (33mer) as previously described [12]. Duplicate filters were hybridized to the 0.7 kb *PvuII/HindIII* fragment of pTrp5.9 (encoding pol II sequences, [12]) at a final stringency of  $1 \times \text{SSC}$  at 50°C. At this stringency this probe also cross-hybridizes to the Trp6 locus [16], encoding Pol III sequences. By selecting those clones which only hybridized to the 33mer, we excluded clones encoding Pol II or Pol III sequences.

### 2.2. DNA sequencing

The isolated clone was mapped by Southern analysis. Genomic fragments of the 11 kb *HindIII* fragment were subcloned into pEMBL vectors [17]. Fragments of the Trp11 locus, containing the 3' half of the gene were isolated using fragment B (fig. 1) as a probe. Progressive unidirectional deletions of the inserted DNA were created using *Bal31* exonuclease. DNA sequence analysis was performed using the dideoxy chain termination method [18] with modifications as described [19]. Both strands were sequenced using 7-deaza dGTP (Boehringer Mannheim) as substitute for dGTP. Sequence analysis was performed with the program described by Queen and Korn ([20]; Microgenie<sup>TM</sup>, Beckman Instruments).

## 3. RESULTS

### 3.1. Identification and isolation of Trp11

All eukaryotic RNA polymerases analysed to date are characterized by a highly conserved eleven

amino acid motif, YNADFDGDEM<sup>N</sup> [2,21]. To identify all putative loci encoding trypanosomal polymerases, we hybridized *HindIII* digested genomic DNA of *T.brucei* with the 33mer, <sup>5</sup>TAC/T AAT GCT GAC/T TTC GAC/T GGT GAC GAA ATG AAT<sup>3</sup>, encoding this motif. This probe hybridized to four fragments, respectively, 4.8, 5.9, 11 and 28 kb in size. We have previously shown that the 4.8 and 5.9 kb *HindIII* fragments encode Pol II sequences [12], while the 28 kb *HindIII* fragment encodes Pol III sequences [16]. This suggests that the remaining locus, Trp11, might encode Pol I. This locus was isolated as a recombinant clone from a genomic library in EMBL 3 phage (see section 2.1 for details), and used for further analysis (fig. 1).

### 3.2. Sequence analysis of Trp11

The nucleotide sequence and the deduced amino acid sequence of the gene encoded by the Trp11 locus are shown in fig. 2. The sequence revealed a single open reading frame of 5052 nt, which is in line with the size of the mRNA, 5 kb, detected in Northern blot experiments [22] with probe A (fig. 1). The predicted molecular mass of the trypanosomal subunit, based on the sequence data, is 185 kDa which is in the size range of other eukaryotic Pol Is [1,23]. The similarity of the deduced amino acid sequence and the presence of the eight homology regions A-H (fig. 2), which are characteristic for eukaryotic RNA polymerases [2,15,16,24], strongly suggest that Trp11 encodes a polymerase subunit. A dot-matrix analysis ([20]; data not shown) revealed that the sequence most closely resembled that of yeast Pol I.

A direct comparison of the trypanosomal Pol I amino acid sequence with that of the Pol I subunit of *Saccharomyces cerevisiae* [21] and *Schizosaccharomyces pombe* [25] reveals several notable features. The 18 amino acid insertion present in the B domain of both yeast subunits is absent in the trypanosomal subunit and is substituted for a single proline residue. Both yeast subunits contain two additional class-specific insertions [16] of approximately 100 mainly hydrophilic residues localized between domain A and B and between G and H [21,25]. Analogous insertions are present in the trypanosomal subunit. A striking feature was the marked difference in length of the amino acid stretches separating the G and H domains. This re-



Fig. 1. Restriction map of the Trp11 locus of *T.brucei*. The coding region of the largest subunit of RNA polymerase I is indicated by a box. The probes used are indicated below the map. Abbreviations of restriction enzyme sites: A, *AvaI*; B, *BamHI*; E, *EcoRI*; H, *HindIII*; S, *SalI*; X, *XbaI*.

-100                      -80                      -60                      -40                      -20  
 CACAACCCATCGCTCTTACAGACTTCCATTCCATGCTTACTTGGGTGACCTGAAGACCCCGCTGTGATGGCGCCAGGACCGCTGCTCTCTGCTTGCATCTGAGGAT  
 1                      20                      40                      60                      80                      100                      120  
 ATGTCGGGATCGCTTTCGTGGAGTTCCACCGGTGCCGCCAGGAGATCGTTCCGACCTTGGCTCCAGTTGTAAAGATGTTGAGAAATCACGCCACATCTACGACACTCGAATG  
 N S G I A F V E V R T R A G Q E D R S A P W R P V V R M G E N N A T F Y D T R M  
 140                      160                      180                      200                      220                      240  
 GGTAACCTCGACCGTAACCGCTTCCACCACAGACATGTCAACGTGCCGACGAGCTTGACAGAAATACGGCAATGAAGTTGTACCGTCACTTTGGAATTTGTGGAAATGCCCGT  
 G N F D A N P F P P Q T C Q T C A A S L T G K Y G N E R C H G N F G F V G W P R  
 260                      280                      300                      320                      340                      360  
 ATCCGCCAGCGAGTCCGCTTCCGACTCCGATCGCTTGTGCTTGAATCCGACCTCGCGATGGATCGAGATCGTTTCCGAGCGAAATGCTTTTGTGCAGATTTGAGCA  
 I R P G S A M S D S D R L V L V L N P H L A N D A D R L F R A K C F F C H K F R A  
 380                      400                      420                      440                      460                      480  
 CCAACATTTGATGTCGAACGATTCGCCAGGCACTGGTTTGGCCGACCATGGTTTACCAGGAGATCGCTGCTATCTTCTGCACAGTGGCAACAGCGGAGGTTCAGACGCTATGCTG  
 P T F D V E R F R Q A L V L A D H G L P G D A L H L L D T V P T A K G H D A N L  
 500                      520                      540                      560                      580                      600  
 AACCAACGCGGCTGCAATGAGCAATCGTCAACGATGTATCCATCCTTCAGTCGTATGTTGATCGTATCTTAAGGACGCGCTAGTGGATGCAGTGAGGAGGTTCGAAACGACGA  
 N H R R M A N E E I V M O V S I L Q S Y V D R I L R Q R A S G C S E E D A K A R  
 620                      640                      660                      680                      700                      720  
 GTGACTATGTCACAGGAGGAGCTGGAGCTTCGCAATGACATTTGCAACATGGCGATTAGCCATTAAGATCAITTCAGTGGTCTTGTAGCCACTGCAGTCTATTTCTCGACCTTC  
 V T H A Q K G T V D V R M D I C H M A I S H L R S F S G P C S H C T A I S P T F  
 740                      760                      780                      800                      820                      840  
 TTGAAGCGCGCGGTATTATTTCTTTTATTCAGGAATCGAATCTTGTGACAAACATTGCCAAGGATTCCTTACCAGCAAGAGGTATCTGAGTGGGAGGTGTCAACCGCTGCAI  
 L K R G G I I F F L F R K S N L V T N I A K G F L T Q E E V S E W E A V N R L H  
 860                      880                      900                      920                      940                      960  
 GGGAGGCTGGAACATATTTGATGGTCTCAATGCTTTTCATATGAAAACCTATTCGAAAAGAGCAAGTATCCTTGGTTTGTGTACCAAACTTTGGTGAACGTCGGTTTC  
 G R T G T Y F D G R Q N L F H M K N L F A K E Q A I L G L L T P M L G E P S V F  
 980                      1000                      1020                      1040                      1060                      1080  
 ACCAAACCAACAGGTTCGCCAGCAAGTGAAGGTACAACTGTCTTTTGGACGCTATTCGTACCGCATTACCCCTGAGGCTTTCGTACGCGTGAGAGTAATGATAATGGT  
 T K T N K V P A S E R Y K L F F L D R I L V P P L P L R L S S G V R V N D G  
 1100                      1120                      1140                      1160                      1180                      1200  
 TTGATATACCGGAGCAACACGCTCTTTCGATATATGGGTTTGTGAGCAGATTGAGTGCITTCACACATGAGTGCTAATCTACAAACGGGCGTAGTTTCATCACTGAT  
 L I I P D E Q T R A L S D I L G F V E O I E C F H T L S A N S T N G R S F I T D  
 1220                      1240                      1260                      1280                      1300                      1320  
 GCGCAGAGGGCTGTGATGAGTGAACCTCCGCAACCTTCAACAAAAGTGGATGAGTTTATGCAGAAATCGTAATAGCTTTGCGAAGAGGAGGAGCTCTTCCGCTGAATATGATG  
 A Q R A V N E S M L R N L O Q K V D E F Y A E I V N S F A K K E G L F R M H M H  
 1340                      1360                      1380                      1400                      1420                      1440  
 GGAAGCGTGTCACTCAGGCTCGCTTTCGGTTATTTCCGCTGATCATTCTGTTGAACCGAAGGAGTCTTTTACCAAGACCACTAGCCGCTTTCGTGCTTCTGAGCAAGTACT  
 G K R V N D A C R S V I S P D F V E P N E V L L P R P L A R A L S F P E D V T  
 1460                      1480                      1500                      1520                      1540                      1560  
 TGTTCGACAGCTCGCTGAATCTGTTGAAGCACTCGCTGTGAACGGCCACGTAAATACCCGGCGCCACACATGAGCTTCGTATGCCAAGGTGAGATCGCTCTGTGAC  
 C F A P A R M N L L K H C V V N G P R K Y P G A T H I E L R N A N G E I R S V D  
 1580                      1600                      1620                      1640                      1660                      1680  
 CTTAATGTACCGCAGCAGCGCGGCGAGCATGCTGCAAGTCTTTTGGATGGCCAGAGTGGTGACCGTAATCGTTTATCGCACATCTTAATGGTGTGCTGCTATATTAAT  
 L N V P E Q T R R Q H A A R F F A M A D S G V T L I V Y R H I L N G D R V I F N  
 1700                      1720                      1740                      1760                      1780                      1800  
 CGCAACCCACACTCCATAAACCGATGATGGGTATCGCTGAAGGTACTTTCAGGGTCCAAAACATTTCGCTTTCACCTGATGATGGTAACTCTTTAAGCAGACTTCGATGGT  
 R D P T L H K P S M H G Y R V K V L S G S K T I R F H Y V N G N S F M A D F D G  
 1820                      1840                      1860                      1880                      1900                      1920  
 GATGAATGAAGCTTCAGTTCCGCAAGCATTGAGACGGCGCCAGGTTGAACCGTGATGGAGCAACATCAACTACCTTGTGCCGAGCTTGGCAGACGATCCGTGGCTTATA  
 D E M N V H V P Q S I E T R A E V E T L M D A N I N Y L V P T S G R P I R G L I  
 1940                      1960                      1980                      2000                      2020                      2040  
 CAGGATCATGTGGCGGAGCGGTTTTGGTAACGTTGGCGGACAGTTCTTGTATCACTCCACCTTGTGCAACTGGTGTACAAATGGCGTGGTCCATACATTGAGGAAACGTTGGCATA  
 Q D H V A A G V L V T L R D K F F O H S T F V Q L V Y N G V G P Y I Q E N V G I  
 2060                      2080                      2100                      2120                      2140                      2160  
 ACTCTTGTGAACCTATTCCTATTCAGCAATCTTATGCTCGGCCAATGGACTGGAACACGCTGATATCTGTGATGGTTGGTTTTCTAGTGGATGTCTCGCCGAGGAGCTGT  
 T L A E L I P I P A I L M P R P M W T G K D L I S V M V R F S S G L S A A S D C  
 2180                      2200                      2220                      2240                      2260                      2280  
 GGACGGGAGATAGAGGGGGAATCACTCAAGGCACTTCAAAATCCAACCCAGCGATTGACAGAAATACCGCGGGTAGCTGGACGCGAGTGGTGGCAATCCGGGCACTGGT  
 G R E I E G G I T L K G T S O I Q P S A F D R I P A G S C D A V R A K S G A V V  
 2300                      2320                      2340                      2360                      2380                      2400  
 GATTCATCTGATGTTTGGCAACAGTGAAGTAAATACCGGATCATGTGTAAGAAGCACTTGGAGCCTTAATATGTCTGCTCCCAACCATGTCTATGAGCTTACGACCATATAGG  
 D S T V M F A N S E L I T G F M C K K Q L G A S H M S A P H V Y E L Y G P R  
 2420                      2440                      2460                      2480                      2500                      2520  
 ACGGACAGTTGTCTGCTCTTTGGCGGTGTCTCTGCTGGCTACGAAAGGAGGCTTTCCCTTGGGATGACGATATGTTCTGTTGATGAAGAGCGGAGTGGAGCTTGT  
 T G Q L F A A F G R V L L L A L R K E G L S L A M D D H F L V D E E R C D L L  
 2540                      2560                      2580                      2600                      2620                      2640  
 AGGAAGCTGTGATGATAGCTTGGATGTTCCAGATGAAGAGGCACTGCTGCACCGATGATTGCAGATTATGCAACAAAGATTGACGAGGAGTTGTTCCGCGAGGCTGCTGGTGGCC  
 R K L D D I A L O V P D E E A T A A P H I A D Y A T K I Q D E F V P Q R N L V P

2660	2680	2700	2720	2740	2760
TTCCGAAGAAATCATCTCTCTCTGATGACTATTTTCAAGTGGGAAAGGAGCAACCTGAAGGCCACACAAATGTCACTGCAAGTTGGCCGACAGCTGTTGATGGAGCTGGCGTGAAGCT					
F P K E H L L L L M T I S G A K G S N L M A T Q M S L Q L G Q L F O G L R V K R					
2780	2800	2820	<b>F</b>	2840	2860
ATGAACCTTCCAAGACGCTTCGTCCTTCTCACTAATGAAAGCGTCCCGCTCTTGGGTTTGCATGGGTTCCCTTCGCTCAGGGATCAGACCGCGGATATACCATTCATGCA					
M N S S K T L P S F F T H E K R A R S F G F A M G S F A S G I R P A E T T I H A					
2900	2920	2940	2960	2980	3000
ATGGCTGGTGGGACGGCTTATTGATACCGCTGTGAAACCTCCGCTTGTGTCACCTGCAACGCTGCCTCATCAAGGCTTCAAAAGCTTGTGGTCCATTGGGACCGTACGGTGGCT					
M A G R D G L I D T A V K T S R S G H L O R C L I K G L E S L V V H M D R T V R					
3020	3040	3060	3080	3100	3120
GACTCCAGCGGCGCTTATCAATTTATGTATGGGGTCAAGGCTTACCGCTGCAAGGCTTCTACGCTTACCGCTGGGAGATGATGAAGGATAACGATAGTGTATGTTTCCAAGCGA					
D S N G S V I Q F M Y G G D G L D P C K A S T L T A W E M M K D N V V D V S K R					
3140	3160	3180	3200	3220	3240
TTCCGAGGAGTGCCTCAGAAAGTGTTCGCGTGGCGAGATGGCGCGGCGAGGGTTGAAGGAGATGGCAACCAAGATGGAAAGCCACGACTGAGGCGGTGCAAAATGCCATATG					
F G G D A S E S V A G A E D G A A A G L K E M R N E D G K P T T E A V Q N A H					
3260	3280	3300	3320	3340	3360
GAGGCAACTATCAGCTATCTCTCCGCGCTCACTCGCAAGAGTTTGTCTGAGTATCTGTGCAAAAAGGCGATTTCGCGCTCTCCGCAAGGTTGCGACCTAGCGCTTGGGAT					
E Q Q L S T Y P L P A S L D K S L S E Y L C K K A D F P L F R K V S T L A R M D					
3380	3400	3420	3440	3460	3480
CGCAACAGCAGCTAAAGAGAGGTTACAGCAAGGGCGCAAGTGGTGGTGGCTTCAAAAACGCTTGCAGACATTACTGCCCGCAGGCGTCTGTGGGCATTGTGCAACCGGGT					
A K Q Q L K E R L Q Q R R Q K V V G A F E K T L A D I T A R R R L V A L C E P G					
3500	3520	3540	<b>G</b>	3560	3580
GAGCGCGTTCGCTCTTGCAGCGCAGGCGTGTGAGCGCTCAACACAGATGACGCTCAACACATTCCACACTGCTGGTTCACTGTGTCCACGCTGACGCGAGGCTATTCTCGACTC					
E P V G L L A Q A A G E P S T Q M T L N T F H T A G S T V S H V T E G I P R L					
3620	3640	3660	3680	3700	3720
CGGAGTTACTGATCTACCGCTCCGTTAATAGCGCGCTTGTGTGTACCGCTTACTAATGCTACAGAGGAAGCAGCAAAAGTATGCCCAAAATGCTTGGTGCAGGTGTCGCCCAAAA					
R E L L I Y A S V N K A A V V V P V T N A T E E D E K V I A K M L R A G V A A K					
3740	3760	3780	3800	3820	3840
CTGACCGAGTGTCTCGGAGGTCACAGATGGGCTGGTGGCGCAAGCGCTTCCAGCTCGATGCAAGCGCACTTAAACACTGGCTTGGTAAGGGCTACCACTATCACTGTGCGCGAGGC					
L T D C L A K V T D G A G G Q S A S S M Q R N L N T G F G K G Y N H V A R K					
3860	3880	3900	3920	3940	3960
CGEACTGGTATGGTATTAGGGTGTCTTCTGTCTTCCGCGCTCGTGTGGAGGAGTTCGCTAAGCGCATGTGTATGTCACTATCGAATCGCAATCTTTTACTGAAGCACTAAG					
R T G M V I T V S F L F S R S C L E E L R K R M C M S P S E N R Q S F T E A L K					
3980	4000	4020	4040	4060	4080
AATGTTTGGCGCTCATCATGAGGTCACTGAGCGCGCTTCCACCGGAGGAAAGCGGTCGCGCTTACGGGCACTTGGAGGATGAAGCGGCGAGTGGCGCGCGCGACAGAGAGA					
N V V R L I M R S L S A V P R E K E S G D G S G H T G G N K G G S G R A D R K R					
4100	4120	4140	4160	4180	4200
AAGCGATCGGCGCGGACGATGGCGGTGCTCTTGGGGTACTTTCGGAGACGAAATATGCCGATGAAGAGGAAACCGATTCTGATGATGGCATGCGAGAGGACGATAGGT					
K R S G P D D G G G P L G G T F G D E I M R I E E G T D S D D G M S E R S S I G					
4220	4240	4260	4280	4300	4320
GGGGTCCGCTGGCTCGGAGGTCTCATCTCTACACTCGGATGGAACCGATACCAAGGGCATACGAGGAAGGATACGGCGCGCGCGCAGCGAGCGAGGGGCTGTGTGAAGTGGCGCG					
G G R A G S E V S S L H S D G T D T R G I A G S D T G G P Q R R R G S V E S G R					
4340	4360	4380	4400	4420	4440
GGTCAAGCTGCTTCAAGTGAAGCGCTGACCGGCTTATATGCCAGAGGAGTGGCTTCGGCGAGGATGCGGAGGATGAGGTGAGATGAGGATCGGAGCGGCGGATGG					
G D D A S D S E A A D P D L Y A R R S G S P A R D A E D G G E M Q D R O G T D W					
4460	4480	4500	4520	4540	4560
GGAGGAACCTGATGCAAGGTGTGGATATGCAACCTCCAGAGATTCAATGAGTTTCAACAGTCAAAATTCGGTGTCTTATTCGCTCCCTTACTGCTGAGCGCGCGCGA					
G G T S M Q G V V G Y D N F P E I H M S F T K S N F G A V I A P L S T A A A A R					
4580	4600	4620	4640	4660	4680
GATGGGTGTGCACTACACGAAGACTTTTTCATAGTTAATGCTGTCTGCGACGGCTCGGAGCTTATCGCGCTTCCGAGCTTGTGGATAATGCTTGGAGGCGGAGAGATG					
D G V V Q L H E D F F I Y N A V L R T A S D V I A V I P D V V D N A L E A Q R M					
4700	4720	4740	4760	4780	4800
CCCTGCTGGTTACTCAATTCGGGTCCCTCACTTTCACCTGACTGAAGATAAAGGATCTGGCGAGCTTGTTCCTCAAGGACCTGGCTCGCAATCGGAAATGTGATGTCTTCTTTC					
P S W L P D F G S L T F T R L K D K G S G D L V F D G P G S T M R N V M S F L S					
4820	4840	4860	4880	4900	4920
CTTTTCACCGTTGGCATAAAATCGATAAATTCACCAAGCTTGTCTCACTGATATCGGGATATGGGAACATACTTCGGCATTGAGTCTGGATACGCTGCGCTTACGATGAGCTCAATA					
L E T V G T X S I M F T K L A P L I F G I M E N T S A L S L D T L R F T M S S I					
4940	4960	4980	<b>H</b>	5000	5020
AACTATTCACCGTACAAATGTCAGCGCGGCGCTTGTGCTCATCGCGGATAGCTTACCCACCGAGGCGGATGGGAGACCTTCAATTCACCTGGTGTCTATCAAGCGCGGAGGC					
N Y S I A T M S T R G T C R S S P I R L P T E G D G R T S I S L V S Y D R A R A					
5060	5080	5100	5120	5140	5160
CACGTGTTTCAGATGACATTTGCTCGTCAAGCGCTGGTGGCAGAGCGGCTGAGTGGGGATATCAACGAGCGGAGCGCACTGTTTCAGATGACATTTGCTCGTCAAGCGCGTGGT					
H C F R *					

Fig. 2. Nucleotide sequence of the Trp11 gene of *T. brucei*, and the corresponding deduced amino acid sequence. The highly conserved homology regions A-H, which are characteristic for all genes encoding the largest subunit of the different eukaryotic RNA polymerases [2,15,16], are boxed. The sequence motif discussed in the text is overlined. The nucleotide sequence presented here has been submitted to the EMBL/GenBank database under the accession number X14399.

gion is approximately 40 amino acid residues longer in the trypanosomal subunit. These amino acids are, however, not clustered. Although this stretch shows a low overall homology between the trypanosomal and the yeast subunits, a single heptapeptide sequence EEG(T/I)DSD was found at, respectively, 178 and 176 amino acid residues after the G domain of the trypanosomal and the *S.cerevisiae* subunits (fig. 2). This motif is not present in the RNA polymerase classes. However, the absence of this motif in *Sc.pombe* Pol I [25] reduces the significance of this motif. Our comparative analysis showed two other interesting differences of the trypanosomal pol I: (i) the absolutely conserved eleven amino acid motif YNADFDGDEM N [2,16,21] contains a single conservative change (FNADFDGDEM N); (ii) the H domain has the lowest degree of homology between identical domains of the individual subunits and is, moreover, enriched in serine and threonine residues.

#### 4. DISCUSSION

##### 4.1. RNA polymerase I-specific domains

Although our knowledge of the role of the individual subunits of the RNA polymerases is still fragmentary, it is clear that the largest subunit has an important functional role in the transcription process [1,2]. A direct comparison of the primary sequence of the trypanosomal (this paper) and yeast Pol Is [21,25] revealed the presence of two additional domains in the N- and C-terminal regions, which are absent in all analyzed Pol IIs [12–15] and Pol IIIs [13,16]. These Pol I-specific domains are highly hydrophilic, suggesting that they might be exposed to the surface and could have a functional role in the RNA polymerase I complex (cf. [21]). The significance of these domains may be in: (i) the direction of the enzyme complex to its nuclear compartment, the nucleolus, or (ii) the interaction with the nucleolar protein nucleolin ([26] and cited references), or (iii) the binding of class-I specific transcription initiation factors [27].

More detailed studies will be necessary to identify the actual role of these Pol I-specific domains in RNA polymerase I transcription.

##### 4.2. Trypanosomal RNA polymerases

The VSG and ESAG genes of *T.brucei* are transcribed in an  $\alpha$ -amanitin resistant fashion [5,8,9].

The exact mechanism of this process and the type of polymerase involved are still unclear (see [10–12]). In order to understand the basis of the  $\alpha$ -amanitin resistant transcription, we have started to analyze the trypanosomal RNA polymerases. For this purpose, we have isolated clones encoding the putative coding regions of these proteins (this paper and [12,16]).

In this report, the structure of the gene encoding the Pol I subunit was determined by DNA sequence analysis. The Pol I structure of *T.brucei* followed the basic scheme typical for all polymerases analyzed so far (fig. 2; cf. [2,15,16]). The major deviations of the trypanosomal Pol I from the identical subunit of the yeasts *S.cerevisiae* and *Sc.pombe* are: the absence of an 18 amino acid insertion in domain B, the length of the amino acid stretch separating domain G and H, and the lack of significant homology of the trypanosomal domain H, which is enriched in serine and threonine residues.

Unfortunately, these observations give no insight into the way these differences might affect trypanosomal RNA polymerase I transcription. As could be expected, they also provide no insight into whether or not RNA polymerase I transcribes the VSG transcription unit. However, the isolation of genomic clones encoding the trypanosomal Pol I–III subunits (this paper and [12,16]), will allow us to prepare specific antibodies by subcloning appropriate fragments into expression vectors and subsequent immunization of rabbits. These antibodies can be used as structural and functional probes in in vitro experiments and might help to identify the RNA polymerase transcribing the VSG genes.

*Acknowledgements:* We wish to thank our colleagues for valuable comments on the manuscript, Mr Klaus Lamberty for artwork and Ms Carmen Müller for photography. This work was supported by the Bundesministerium für Forschung und Technologie grant 0318885 and a research award of the Erna and Victor Hasselblad Foundation to Albert W.C.A. Cornelissen.

#### REFERENCES

- [1] Sentenac, A. (1985) CRC Crit. Rev. Biochem. 18, 31–90.
- [2] Cornelissen, A.W.C.A., Evers, R. and Köck, J. (1988) Oxford Surveys Euk. Gen. 5, 91–131.
- [3] Wieland, T. and Faulstich, H. (1978) CRC Crit. Rev. Biochem. 5, 185–260.
- [4] Cully, D.F., Ip, H.S. and Cross, G.A.M. (1985) Cell 42, 173–182.

- [5] Kooter, J.M., Van der Spek, H.J., Wagter, R., D'Oliveira, C.E., Van der Hoeven, F., Johnson, P.J. and Borst, P. (1987) *Cell* 51, 261–272.
- [6] Johnson, P.J., Kooter, J.M. and Borst, P. (1987) *Cell* 51, 273–281.
- [7] Gibbs, C.P. and Cross, G.A.M. (1988) *Mol. Biochem. Parasitol.* 28, 197–206.
- [8] Kooter, J.M. and Borst, P. (1984) *Nucleic Acids Res.* 12, 9457–9472.
- [9] Laird, P.W., Kooter, J.M. and Borst, P. (1985) *Nucleic Acids Res.* 13, 4253–4266.
- [10] Shea, C., Lee, M.G.-S. and Van der Ploeg, L.H.T. (1987) *Cell* 50, 603–612.
- [11] Alexandre, S., Guyaux, M., Murphy, N.B., Coquelet, H., Pays, A., Steinert, M. and Pays, E. (1988) *Mol. Cell. Biol.* 8, 2367–2378.
- [12] Evers, R., Hammer, A., Köck, J., Jess, W., Borst, P., Mémet, S. and Cornelissen, A.W.C.A. (1989) *Cell* 56, in press.
- [13] Allison, L.A., Moyle, M., Shales, M. and Ingles, C.J. (1985) *Cell* 42, 599–610.
- [14] Ahearn, J.M., Bartolomei, M.S., West, M.L., Cisek, L.J. and Corden, J.L. (1987) *J. Biol. Chem.* 262, 10695–10705.
- [15] Jokerst, R.S., Weeks, J.R., Zehring, W.A. and Greenleaf, A.L. (1989) *Mol. Gen. Genet.* 215, 266–275.
- [16] Köck, J., Evers, R. and Cornelissen, A.W.C.A. (1988) *Nucleic Acids Res.* 16, 8753–8772.
- [17] Dente, L., Cesareni, G. and Cortese, R. (1983) *Nucleic Acids Res.* 11, 1645–1655.
- [18] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463–5467.
- [19] Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3963–3965.
- [20] Queen, C. and Korn, L.J. (1984) *Nucleic Acids Res.* 12, 581–599.
- [21] Mémet, S., Gouy, M., Marck, C., Sentenac, A. and Buhler, J.-M. (1988) *J. Biol. Chem.* 263, 2830–2839.
- [22] Cornelissen, A.W.C.A., Evers, R., Grondal, E., Hammer, A., Jess, W. and Köck, J. (1989) *Nova Acta Leopoldina*, in press.
- [23] Paule, M.R. (1981) *Trends Biochem. Sci.* 5, 128–131.
- [24] Mémet, S., Saurin, W. and Sentenac, A. (1988) *J. Biol. Chem.* 263, 10048–10051.
- [25] Hirano, T., Konoha, G., Toda, T. and Yanagida, M. (1989) *J. Cell Biol.* 108, 243–253.
- [26] Borer, R.A., Lehner, C.F., Eppenberger, H.M. and Nigg, E.A. (1989) *Cell* 56, 379–390.
- [27] Kownin, P., Batemann, E. and Paule, M.R. (1987) *Cell* 50, 693–699.