Volume 9 Number 18 1981 Nucleic Acids Research

Homologous 3'-terminal regions of mRNAs for surface antigens of different antigenic variants of Trypanosoma brucei

Hemanta K. Majumder, John C.Boothroyd* and Hans Weber

Institut für Molekularbiologie I, Universität Zürich, 8093 Zürich, Switzerland, and *Wellcome Research Laboratories, Department of Immunochemistry, Langley Court, Beckenham, Kent BR3 3BS, UK

Received 16 June 1981

ABSTRACT

Sequences corresponding to the complete 3'-terminal regions of the messenger RNAs for three different Variant Surface Glycoproteins of Trypanosoma brucei were determined on complementary DNA inserts cloned in recombinant plasmids. The three sequences show 80-130 base pair long segments of strong (70-80%) homology at the 3' ends, whereas the regions upstream from the last 130-140 base pairs contain no significant homology. The signal AAUAAA, present near the 3' ends of almost all known polyadenylated mRNAs of eukaryotes, does not occur in the 3'-terminal sequences of these three variants.

INTRODUCTION

The mechanism generating the diversity of the Variant Surface Glycoproteins (VSG) of African trypanosomes has recently attracted considerable interest (1-3). Recombinant plasmids containing inserts of DNA complementary to mRNA for the VSGs of several cloned trypanosome variants (4) were important tools for these studies. Nucleotide sequence analysis of several of these inserts revealed that the coding regions of the VSGs include the codons for hydrophobic amino-terminal leader and carboxy-terminal tail peptides which are lacking on the mature VSGs (5,6). Two inserts of antigenic variants (types 117 and 221) showed no sequence relationship, except for their 3'-terminal regions, which were highly homologous. In addition, it became clear that these, like most of the inserts initially characterized, were missing nucleotides not only from the 5' end of the mRNA (as expected from the cloning procedure) but also from the 3' terminus (as evidenced by the lack of poly(dA:dT) regions). In a search through the original plasmid collection we now have isolated VSG-specific inserts of three different

antigenic types carrying intact 3' ends. Nucleotide sequence analysis showed that all three contain 80 to 130 nucleotide long 3'-terminal regions of high (70-80%) homology and that they lack the putative poly(A) addition signal AATAAA.

MATERIALS AND METHODS

The construction and selection of recombinant plasmids containing VSG-specific cDNA inserts in the PstI site of the vector pBR322 was described by Hoeijmakers et al. (4). The collection of clones carrying such plasmids was obtained from Professor Charles Weissmann. Restriction mapping and Southern hybridization (7) with 3'-terminal probes were used to identify inserts of a given type extending furthest to the 3' end of the mRNA. Since the recombinant clones picked originally on the basis of differential hybridization did not contain inserts of variant types 117 and 118 carrying intact 3' ends, the collections were screened once more by colony hybridization (4), using labeled fragments from incomplete inserts as probes. Nucleotide sequence determinations were carried out according to Maxam and Gilbert (8). Similarities of sequences were calculated by the procedure described by van Ooyen et al. (9).

RESULTS AND DISCUSSION

The VSG-specific cDNA inserts containing complete 3' termini described here are from three different antigenic variants of T.brucei: inserts TcV 117.E and TcV 117.W are from variant 117, TcV 118-CII.4 and TcV 118-CII.10 from variant 118, and insert TcV 221.1 from variant 221. These three variants arose from a single cloned stock of T.brucei. Fig. 1 shows restriction maps of the inserts, along with the maps of some inserts carrying incomplete 3' termini which were isolated earlier and were also used in the present work. Surprisingly, inserts TcV 117.E and TcV 117.W seem to extend to exactly the same point at their 5' ends, within the accuracy of the mapping method using polyacrylamide gel electrophoresis. However, they cannot be descendants from the same cloning event since their nucleotide sequences are not exactly identical (see below). Since on a statistical basis such a coincidence is extremely improbable, this 5'-

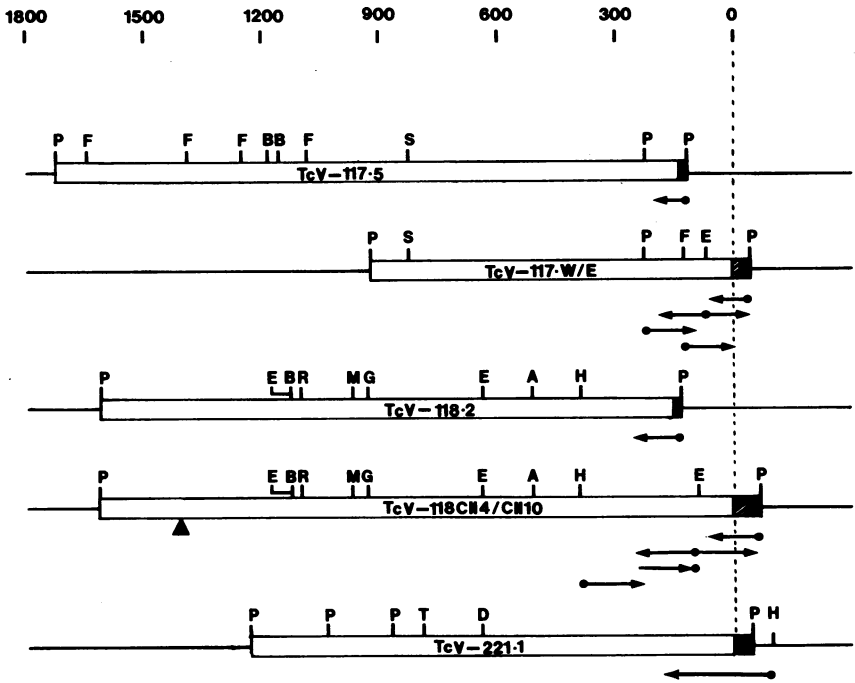

Figure 1. Physical maps of several cDNA inserts of VSG variant types 117, 118 and 221, showing restriction endonuclease cleavage sites and regions sequenced. The boxed regions show the lengths of the inserts; poly(dC:dG) tails are indicated as black and poly(dA:dT) blocks as hatched segments, the vector pBR322 is drawn as a single line. The inserts were aligned at the beginning of the poly(dA:dT) block; if this was absent, by comparison of the restriction maps. Distances are in base pairs from the beginning of the poly(dA:dT) block. The 5' end of TcV 118-CII.10 is indicated (▲). Restriction sites are coded as follows: A, AluI; B, BglI; D, HindIII; E, BspI(HaeIII); F, HinfI; G, BglII; H, HpaII; M, BamHl; P, PstI; R, EcoRI; S, SalI; T, TaqI. A number of sites for restriction enzymes AluI, BspI, HinfI, HpaII and TaqI are omitted from the figure. Below the inserts are shown the sites which were 5' or 3' end labeled (●) and the regions where sequences could be read unambiguously (⟶, indicating 5' to 3' direction).

terminal site might represent a preferential break-off point during the cloning procedure, e.g. in the reverse transcriptase reaction.

Colony hybridization experiments using 3'-terminal labeled probes of inserts TcV 117.E or TcV 118-CII.4 revealed weak but

significant cross-hybridization with colonies of the other two variant types. Similarly, Southern hybridization using restriction digests of plasmids of the three types blotted on nitrocellulose filters indicated that their crosshybridization was limited to the 3'-terminal region of the inserts (results not shown).

The labeling sites and the strategies used for determining the nucleotide sequences near the 3' ends of the inserts are indicated on the restriction maps of Fig. 1 by dots and arrows. The sequences obtained are presented in Fig. 2; most of those shown were determined on both DNA strands at least once. TcV 221.1 was sequenced twice on the same strand; its sequence was found to be unambiguous and in complete agreement (in the region of overlap) with that of TcV 221.12 (6). A segment spanning nucleotides 20 to 87 of TcV 118-CII.4 and TcV 118-CII.10 was sequenced on the plus strand only; both clones gave completely identical and unambiguous results. All sequences shown exhibit a poly(dC) tract of 20-30 nucleotides length at the very 3' end resulting from the homopolymer tails by which the inserts were integrated into the vector DNA. This region is preceded by a tract of 15-50 A residues which originates from the poly(A) tail of the mRNA, which proves that these inserts contain complete 3' termini. The (heteropolymeric) sequences of inserts TcV 117.E/W and TcV 221.1 confirm those of TcV 117.8 and TcV 221.12 presented in earlier work (5,6) and indicate that the latter inserts were actually missing only a very small number of nucleotides (probably 3 to 8) at their 3' ends. By aligning the sequences of the three variant types 117, 118 and 221 as shown in Fig. 2 it becomes evident that they contain regions of strong homology preceding the poly(dA) tract. Comparing the variants 117 and 118, these segments have a length of 133 and 138 base pairs (bp), respectively, with similarity values of 78% and 79% (calculated according to ref. 9, with N=4). Upstream of this region, the similarity drops quite abruptly to 33%. (The value for random sequences is about 31% (9)). The pair 117 and 221 shows, in segments of 80 and 94 bp, respectively, similarities of 68% and 72%; for the pair 118 and 221 the corresponding values are 85/94 bp and 68/71%. A significant similarity (42 to 48%) between 221 and

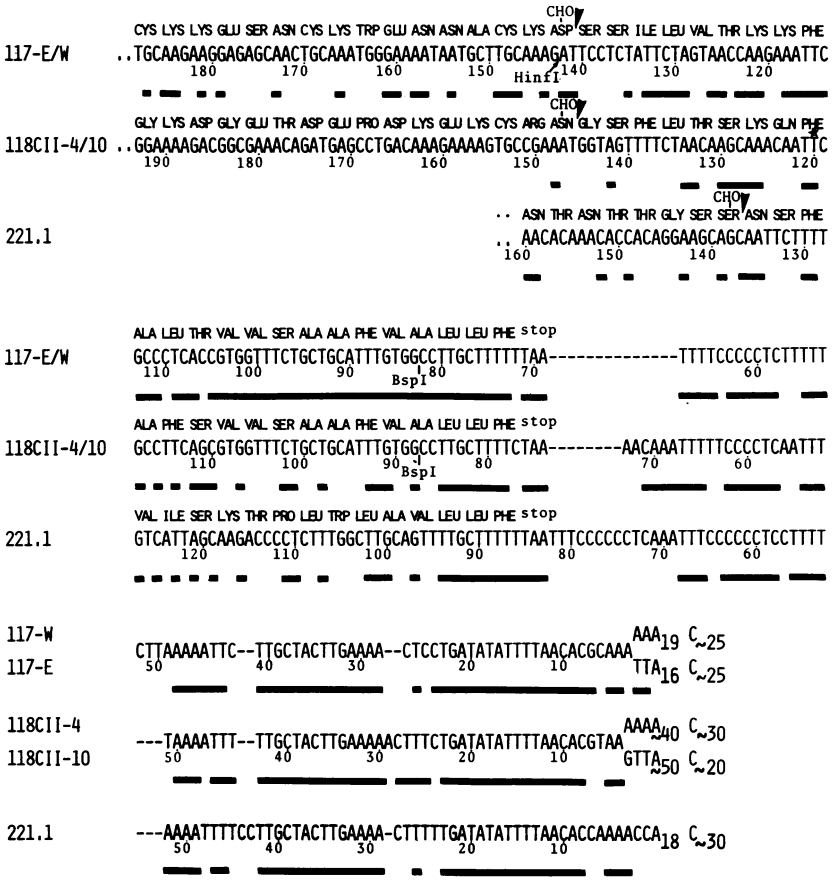

Figure 2. Nucleotide sequences of the complete 3'-terminal regions of cDNA inserts of VSG variant types 117, 118 and 221. Numbering is backwards and starts from the last nucleotide of heteropolymeric sequence. Sequences are aligned for maximal homology, however avoiding an excessive number of gaps (N = 4, ref. 9). Black bars between two sequences indicate homology between the two; the black bars below the variant 221 sequence indicate homology with variant 117. Above the nucleotide sequences the coded amino acids are indicated. The carboxy-terminal residues of the mature isolated glycoproteins are followed by (▼). Glycosylated amino acids are marked by (CHO). (For variant 118 these assignements are only tentative). (*) indicates a nucleotide whose presence could be detected on the sequence gels of only one of the two strands, probably due to a "band compression" effect.

the other two variants is observed for an additional 48 bp upstream, approximately to the point where variants 117 and 118 diverge. Further upstream no significant similarity is evident, but the available sequence seems too short to allow a meaningful statement. The percentage of nucleotides common to all three variant types 117, 118 and 221 in 3'-terminal segments of 80, 85 and 94 bp, respectively, is 78, 73 and 66% (neglecting any "penalty" for the four gap sites), whereas in the upstream 48 bp long segment this value is 42%.

The 3'-non-coding regions of the three variant types are unusual in that they lack the otherwise ubiquitous sequence AATAAA (10,11) or a variation thereof (12) preceding the poly(A)-addition site by about 20 bp. In closer proximity to that site, however, they all contain the sequence TTTTAACACG, which is reminiscent of the "consensus" sequence TTTTCACTGC in the analogous position in several other mRNAs (11). Both in the case of the type 117 and the type 118 variants, the sequences of the two individual clones of the same type are completely identical as far as determined, with the somewhat surprising exception that the position of the poly(dA) block in inserts TcV 117.E and TcV 118-CII.10 is shifted 5 bp downstream relative to the homologous sequences of the clones TcV 117.W and TcV 118-CII.4. The heterogeneity could be due to imprecision in the selection of the polyadenylation site. Recently a similar heterogeneity was described in SV40 mutants which have small deletions between the (intact) AATAAA site and the 3' end of the transcript (13). This result suggests that it is probably not the absence of the AATAAA sequence *per se* which is responsible for this type of heterogeneity in the VSG cDNAs. We cannot exclude the possibility that it is an artefact of the cDNA synthesis (the oligo(dT) primer could have spanned the junction of the transcribed and poly(A) portions of the mRNA). A less likely but still possible explanation is that these two forms reflect allelic differences.

The homology among the three variant types extends into the coding region. The reading frames for the variants 117 and 221 have been determined previously by comparison with peptide sequence data (5,6). By allowing appropriate gaps or insertions the sequences of the three variants can be realigned at the TAA ter-

mination codon corresponding to the known reading frame. Preceding this, all three variants share a 8-9 bp block of common sequence leading to the homologous carboxy-terminal peptide sequence Leu-Leu-Phe, but further upstream the peptide sequence of variant 221 diverges completely (although it remains hydrophobic and the nucleotide sequence retains some similarity). On the other hand, the longer region of strong homology between variants 117 and 118 results in carboxy-terminal peptide sequences in which 14 out of 17 amino acids are identical.

The carboxy terminus of the mature VSGs is thought to arise by a processing step, in which a hydrophobic oligopeptide is cleaved off (5,6). The glycosylated amino acids at the mature termini of variants 117 (Asp) and 221 (Ser) are marked by arrowheads in Fig. 2. Tryptic cleavage of mature variant 118 protein yields a glycosylated Asx residue as carboxy-terminal product (A.A. Holder and G.A.M. Cross, personal communication). Considering the sequence specificity for asparagine glycosylation (14), the partial amino acid sequence in this region can be predicted as -Lys/Arg-Asn(CHO)-Xxx-Ser/Thr-. A sequence -Arg-Asn-Gly-Ser- is in fact coded by the probable reading frame of the variant 118 sequence; it is located outside the region of homology with variant 117 but the position of the carboxy-terminal Asn residue of variant 118 corresponds exactly to that of the carboxy-terminal Asp(CHO) residue of variant 117. VSG 118 is unique among the six VSGs of T.brucei which have been examined in that its C-terminal glycopeptide shows no immunological cross-reactivity with the corresponding, completely cross-reactive glycopeptides of the other five (15; A.A. Holder, personal communication). The finding here that it is the only one of the five with a glycosylated C-terminal Asn (vs. Asp or Ser) may in part explain this exception.

From the results presented here it seems likely that the 3'-terminal homology of VSG mRNAs is a general phenomenon. This phenomenon is probably related to the unusual way in which the expression of VSG genes is controlled. Activation of a VSG gene is accompanied by its duplication and the transfer of the copy

to an "expression site" (1, 16). This "expression-linked copy" (ELC) is the one used for mRNA synthesis (A. Bernards and J.C. Boothroyd, unpublished). In the case of the 117 and 118 genes it has recently been shown that the ELC gene differs from the basic copy gene from which it originates at the 3' end (6, A. Bernards and J.C. Boothroyd, unpublished). The 3'-terminal sequence of the basic copy 117 gene differs from the corresponding cDNA sequence, but has about 70-80% homology with 117 cDNA (A. Bernards and J.C. Boothroyd, unpublished). It seems likely, therefore, that the transposition which gives rise to an active ELC gene involves a cross-over between the 3' end of the basic copy gene and sequences in the "expression-site" providing the gene with a new 3' end. The differences between the 3' ends of the cDNAs from the different variants analyzed here could either come from the presence of multiple "expression sites", from sloppy recombination in the joining region or from a combination of both.

ACKNOWLEDGEMENTS

We thank Professor Charles Weissmann for the collection of trypanosome cDNA clones as well as for advice and support. We are grateful to Mr. A. Bernards, Professor Piet Borst, Dr. George A.M. Cross and Dr. A.A. Holder for discussions and for communicating results prior to publication. We thank Dr. Peter Dierks and Mr. Jürg Schmid for several gifts of $[\gamma\text{-}^{32}P]$ ATP and Ms. C.A. Paynter for expert technical assistance. Supported by the Schweizerische Nationalfonds and the Kanton of Zürich.

REFERENCES

1. Borst, P., Frasch, A.C.C., Bernards, A., Van der Ploeg, L.H. T., Hoeijmakers, J.H.J., Arnberg, A.C. and Cross, G.A.M. (1981) Cold Spring Harbor Symp. Quant. Biol., 45, in press.
2. Williams, R.O., Young, J.R. and Majiwa, P.A.O. (1979) Nature 282, 847-849.
3. Pays, E., Van Meirvenne, N., LeRay, D. and Steinert, M. (1981) Proc. Natl. Acad. Sci. USA 78, 2673-2677.
4. Hoeijmakers, J.H.J., Borst, P., Van den Burg, J., Weissmann, C. and Cross, G.A.M. (1980) Gene 8, 391-417.
5. Boothroyd, J.C., Cross, G.A.M., Hoeijmakers, J.H.J. and Borst, P. (1980) Nature 288, 624-626.

6. Boothroyd, J.C., Paynter, C.A., Cross, G.A.M., Bernards, A. and Borst, P. (1981) Nucl. Acids Res. 9,4735-4743.
7. Southern, E.M. (1975) J. Mol. Biol. 98, 503-517.
8. Maxam, A.M. and Gilbert, W. (1980) in Methods in Enzymology (Grossman, L. and Moldave, K., eds.) 65, 499-560, Academic Press.
9. van Ooyen, A., van den Berg, J., Mantei, N. and Weissmann, C. (1979) Science 206, 337-344.
10. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
11. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucl. Acids Res. 8, 127-142.
12. Hobart, P., Crawford, R., Shen, L., Pictet, R. and Rutter, W.J. (1980) Nature 288, 137-141.
13. Fitzgerald, M. and Shenk, T. (1981) Cell 24, 251-260.
14. Marshall, R.D. (1972) Annu. Rev. Biochem. 41, 673-702.
15. Cross, G.A.M. (1979) Nature 277, 310-312.
16. Hoeijmakers, J.H.J., Frasch, A.C.C., Bernards, A., Borst, P. and Cross, G.A.M. (1980) Nature 284, 78-80.